Spatial patterns and the socioeconomic determinants of COVID-19 infections in Ottawa, Canada.

A thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the degree of Master of Science in Geography.

Department of Geography, Environment and Geomatics

University of Ottawa

uOttawa

Supervisor:

Dr. Michael Sawada

Committee members:

Dr. Konrad Gajewski

Dr. Anders Knudby

Table of Contents

List of Figures

List of tables

Acknowledgment

I am deeply grateful to my supervisor, Dr. Michael Sawada, for their exceptional guidance, unwavering support, and invaluable contributions throughout this research journey. Their expertise, dedication, and feedback have played a pivotal role in shaping the trajectory and quality of this thesis. In addition to serving as an exceptional supervisor, Dr. Michael Sawada has also been a remarkable research collaborator, meticulous proofreader, and a supportive faculty member who has become a trusted friend. Their valuable insights and collaborative efforts across various aspects of the study have significantly enhanced its depth. Dr. Sawada's commitment to excellence and their continuous push to exceed boundaries to deliver exceptional work have been a true inspiration. I consider myself truly fortunate to have the privilege of working under their supervision. Additionally, I am deeply grateful for the immense support from my parents, Mohamed, and Nihel, who cheered me up in moments of weakness and doubt. My sister, Sahar, and her husband, Bilel, who have always been a reliable and supportive backbone to my struggles. I am also thankful for my friends, Ezzeddine, and Youssef, who never failed to put a smile on my face and provided unwavering encouragement and motivation during moments of doubt and lack of motivation, keeping me focused and determined throughout this academic journey. Finally, I would like to extend my deep gratitude to the Ottawa Neighborhood Study for storing and making the data used in this analysis publicly available, as it significantly facilitated my research process.

Abstract

This study uncovered the pattern and spatial relationships between socio-economic factors and aggregated COVID-19 rates in Ottawa, Canada, from July 2020 to December 2021 at the neighbourhood scale. Both top-down and bottom-up data mining approaches were used to predict COVID-19 rates. The top-down approach employed ordinary least squares regression (OLS), spatial error model (SEM), geographically weighted regression (GWR) and multi-scale geographically weighted regression (MGWR). Model intercomparison was also undertaken. The pattern of COVID-19 in Ottawa exhibited a significant moderately positive spatial structure among neighbourhoods (Moran's $I = 0.39$; $p = 0.0001$). Local Moran's analysis identified areas of low and high COVID-19 clustering, interspersed with cold spots. The OLS model used determinants based on a literature review. Determinants were tested for normality using the Shapiro-Wilks test with those that failed the test had transformatoins to normality applied. Next, an OLS-based backward stepwise approach was used to select the optimal set of determinants based on goodness of fit, selecting the model with the lowest Akaike Information Criterion (AIC). The percentage of people who take public transit to work, percentage of people with no high school diploma, percentage of people over 65 years old, and percentage of people with a Bachelor level degree or above comprised the final set of determinants. A SEM model was created to account for residual spatial autocorrelation in the OLS model's residuals and yielded an adjusted $R^2 = 0.63$. Based on the SEM, a one-unit increase in the square root of the percentage of people with a bachelor's degree or above was associated with a 3.2% increase in COVID-19 rates, while the same unit increase in the square root of the percentage of people with no high school diploma was associated with a 10.6% increase in COVID-19 rates. Conversely, a one percent increase in the percentage of people aged 65 and older was linked to a 34.6% decrease in COVID-19 rates. To examine local variations in the relationships between the determinants and COVID-19, a MGWR with a Bisquare kernel and an adaptive bandwidth was used to improve upon the overall explained variance of the SEM model. The residuals of the MGWR model exhibited no significant spatial autocorrelation (Moran's $I = -0.04$; $p = 0.62$) and residuals were approximately normal ($W = 0.98$; $p > 0.25$). The MGWR model yielded an adjusted $R^2 = 0.75$. Taking a data mining and bottom-up approach, an optimized Random Forest model provided a very different set of determinants as important when compared to the top-down regression approaches and accounted for 47.34% of the COVID-19 variance.

Chapter 1: Introduction, Contributions and Literature Review.

1. Introduction

COVID-19, caused by the SARS-CoV-2 virus, was declared a Public Health Emergency of International Concern and a pandemic by the World Health Organization in January and March 2020. By late 2020 and early 2021, vaccines for the disease had become available to the public in many countries. According to Ottawa Public Health officials, the first confirmed case of COVID-19 in Ottawa was announced on March 11, 2020. Then, it was only a matter of time before COVID-19 spread across all of Ottawa. Many parts of the world recognized the need for an effective, systematic way of predicting COVID-19 rates. Geographic epidemiological analyses are an approach to uncover the nature of disease patterns that can be used to support solutions for preventing disease spread.

This study examined the efficacy of predictive models of COVID-19 rates in Ottawa that are built top-down and bottom-up. Top-down modelling requires inductive-reasoning, specifically, examining individual case studies from the learned literature and deriving a set of general health-determinants that are believed to be strongly associated with COVD-19 rates. Once identified, these determinants are used to create statistical models that predict COVID-19 rates. Alternatively, bottom-up or deductive reasoning is what is often used in data mining and machine learning. In the deductive approach machine learning methods like Random Forest are used to find the best set of predictors within a broad set of health-determinants. Both approaches have merit, and this study examines and compares each.

To achieve the intermodel comparisons and to understand COVID-19 patterns in general, the following research questions are posed:

1. What is the spatial structure of COVID-19 in Ottawa and where are significant clusters and outliers of COVID-19 rates in Ottawa?
2. Using a top-down approach to modelling COVID-19 rates based on literature-derived socioeconomic health determinants, can local and global regression modelling accurately predict the pattern of COVID-19 cumulative rates in Ottawa?
3. Would a bottom-up Random Forest model find the same health determinants as important when compared to the regression modelling? Would RF provide a more accurate model?

Ultimately, 1 to 3 informs which socioeconomic determinants contribute the most to the cumulative pattern of COVID-19 in Ottawa. In this intermodel comparison, this thesis aims to provide insights that could assist public health authorities in better allocating resources in future pandemics or in modelling disease rates through the use of aspatial and spatial modelling techniques.

2. Objectives:

The first research question was addressed by identifying and characterizing the pattern of COVID-19 infection rates and their spatial structure. At this stage it was essential to identify neighbourhoods with a significantly higher or lower infection rates in Ottawa, so that their commonalities could be generalized. The overall spatial structure of COVID-19

rates was assessed using Global Moran's I (Moran 1950). This measure to determines whether the global pattern of COVID-19 in Ottawa exhibits self-similarity or self-dissimilarity. If, for example, the overall pattern is positively spatially autocorrelated, then this means that nearby neighbourhoods have similar COVID rates and that alludes to nearby neighbourhoods also having similar socioeconomic structures. Next, to decompose the global measure and identify which neighbourhoods were themselves statistically significantly different or similar to their neighbours, a Local Indicator of Spatial Autocorrelation (LISA) was used, namely local Moran's I (Anselin 2010). The presence of positive spatial autocorrelation locally (high-high or low-low rates) indicates the possibility of COVID-19 clusters, whereas local negative spatial autocorrelation (a.k.a, hot spots, or cold spots, respectively) indicates neighbourhoods that significantly deviate from the global positive spatial autocorrelation, and these are spatial outliers called hot or cold spots. It becomes critical to examine why those local outliers occur - to the extent the data allows.

The second research question of this study was to identify and investigate the influence of socioeconomic determinants on the pattern of COVID-19 rates in Ottawa, Canada. To address that research question, health determinants were identified within the learned literature based on studies that also assessed patterns of COVID-19 rates. Next, ordinary least squares (OLS) regression was employed to assess the relationship between COVID-19 rates and health determinants at the Ottawa neighbourhood scale. The OLS modelling required spatial techniques to account for spatial dependence in OLS residuals. A Spatial Error Model (SEM) was used for that purpose. Furthermore, a multi-scale geographically weighted regression (MGWR) was applied to examine the local variation between COVID-19 rates and the socioeconomic determinants as measured by the beta parameters in a multivariate model.

For the third question, from a large pool of socioeconomic variables provided by the Ottawa Neighbourhood Study (ONS) (Ottawa Neighborhood Study n.d.) that included those from the learned literature search, a Random Forest (RF) analysis was used to identify the most important determinants associated with COVID-19 rates. The RF analysis also served as a means of comparison between machine-learning selected determinants with those that were literature-derived and used in the OLS and SEM models.

Finally, both model projections/predications were compared using differenced maps as well as Lee's L (Lee 2004) measure of bi-variate spatial correlation.

This study sheds light on the impact of COVID-19 on specific neighborhoods in Ottawa and identifies areas that are spatial outliers and require more explanation with the ultimate goal of trying to uncover the driving mechanisms or reasons for their existence. That can aid in future pandemics or modelling endeavours if the mechanisms associated with the elevated or lowered rates are studied. The findings of this study can be used to inform public health policy and resource allocation with the aim of addressing the underlying socioeconomic determinants that are associated with the variation in COVID-19 rates in certain neighborhoods. By recognizing and addressing these underlying determinants, this study promotes equitable health outcomes and ultimately proposes a set of tools that could help reduce the impact of future pandemics.

3. Literature Review

Spatial methods for disease mapping and disease clustering, at the ecological level of analysis (spatial polygon units), are among the most predominant approaches in geographic epidemiology (Rezaeian et al. 2007). In this study, the variables that are associated with COVID-19 rates are referred to as determinants. Disease mapping and spatial clustering measures have successfully been used in the context of other communicable diseases, such as cholera (Ali et al. 2006), dengue fever (Banu et al. 2012), as well as COVID-19 (Darques et al. 2022). Spatial methods can explain patterns of disease spread and help manage medical resource allocation during a pandemic or local epidemic.

Researchers have investigated dengue fever (DF) transmission in Bangladesh to identify high-risk areas using spatiotemporal methods based on using a Poisson regression model (Banu et al. 2012) using the SatScan software (Kulldorff 1997; Kulldorf, M 2023) . SatScan software employs a Poisson regression model to examine the expected number of cases in an area (based on population size and other factors) and compares that to the observed number of cases to determine if a significant cluster exists (Kulldorf, M 2023). The health determinants used in the model included the aggregated monthly number of Dengue fever (DF) cases and deaths as the dependent variable, population, and the spatial coordinates for each district. Upon analyzing the space-time DF distribution and clusters, they identified the peak transmission time intervals and areas most vulnerable to DF. Their study highlighted that socio-demographic changes, climatic factors, and vector control measures influence DF clusters over time (Banu et al. 2012). The study notes limitations such as the use of aggregated district-level data, which may miss local clusters, potential underreporting in surveillance data, and that the identification of clusters without exploring potential risk factors in depth is a weakness of their approach.

Siljander et al. (2022) used COVID-19 infection rates at the postal code level to identify COVID-19 spatial variability and socioeconomic data to uncover the determinants of COVID-19 patterns. The dependent variable was the number of new COVID-19 infections per 100,000 residents in a postal code area during the previous 14 days. The determinants considered were the number of non-native residents, educational attainment at the elementary level, median income, proportion of people belonging to the lowest income bracket, households categorized as having the lowest income, unemployment rate, and number of retired individuals. They used Global and Local Moran's I to measure the spatial dependence of the pattern COVID-19 rates and to identify hot/cold spots within the global pattern (Siljander et al. 2022). Ordinary Least Squares (OLS), Geographically Weighted Regression (GWR), and Multiscale Geographically Weighted Regression (MGWR) regression methods were used to determine which of their sociodemographic determinants best explained COVID-19 spatial variability (refer to methods for an explanation of the OLS, GWR and MGWR techniques). To determine the best set of socioeconomic determinants that explain COVID-19 infection rates, they used the Exploratory Regression tool in ArcGIS (ESRI 2023a) and chosed the model with the lowest Akaike Information Criterion (AIC) value and the highest adjusted $R^2 = 0.401$, they do not however present the information for for all of the the top-*n* models. Because the exploratory regression tool in ArcGIS tries all possible combinations of determinants, the

best model may not have significantly different AIC values compared to say the *n*-1th model. Nevertheless, the results showed that COVID-19 infection rates were positively spatially autocorrelated, indicating that the distribution of rates was not random. Income, foreign citizens, and education level were found to be the significant sociodemographic determinants. The GWR model, based on the same variables, provided the best model with an ($R^2 = 0.453$) and the MGWR model explaining less at ($R^2 = 0.436$). However, they ran the same models for different time periods during the pandemic and found the MGWR model preformed well with its highest R^2 reaching to 0.61, slightly outpreforming the other two methods for the same time-periods. Relying solely on previous studies in the area, Siljander et al. (2022) used only seven sociodemographic determinants to avoid multicollinearity; however, they did not check variance inflation factor (VIF) scores to determine issues with multicollinearity. Morover, one large weakness of using the all combinations regression model is the issue of multiple testing which they did not address. While not unique to this study, the modifiable areal unit problem (MAUP) (A S Fotheringham and Wong 1991; M.-P. Parenteau and Sawada 2011) restricts the effects of the determinants, their importance, significance and interpretation to the postal code scale. Since vaccinations in Helsinki only began in January 2021, the study could not account for the rapid accumulation of population immunity and this limitation is true of the current thesis as well.

Tang, Vieira, and Shearer (2022) conducted a study at the census tract level in Orange County, California. Their multivariate Poisson regression model furnished a positive relative risk for percent of minority, percent working in service industries, average household size, and percent aged 65 and decreased risk for median household income. Cases with missing geographic coordinates, incarcerated individuals, homeless individuals, and residents of long-term care facilities were excluded from the analysis to focus on community risks. COVID-19 analyses frequently exclude these groups due to the accelerated disease contagion observed among institutionalized individuals. This acceleration is inconsistent with the broader population in the vicinity of these institutions. Including them in the modeling would introduce bias to the model parameters, potentially inflating the relative risk when it should remain unchanged. This study successfully provided relative risk for COVID-19 as a function of socioeconomic determinants. The authors noted that this study may have underestimated the risk of COVID-19 in California, because they only used data from individuals who reported a positive PCR test. And all other studies had no information on the asymptomatic cases (Tang, Vieira, and Shear Mansour er 2022).

Mansour et al. (2021) investigated the sociodemographic determinants of COVID-19 incidence rates using global and local regression models to find which determinants explain the variation in COVID-19 incidence rates and how such relationships varied across the study area. The candidate determinants were population density, number of hospital beds, population over the age of 65, diabetes rate, immigration, crude death rate, and number of physicians and nurses. They then used stepwise forward regression analysis to eliminate non-significant determinants. VIF values were also used to test for collinearity between the model determinants. Global regression methods, such as Ordinary Least Squares (OLS) ($R^2 = 0.58$), Spatial Lag Model (SLM) ($R^2 = 0.62$), and Spatial Error Model (SEM) ($R^2 = 0.65$). The OLS model showed that increasing population over the age of 65 years was associated with increasing COVID-19, followed by population density and

diabetes rate showing the same relations. Multiscale Geographically Weighted Regression (MGWR) demonstrated superior overall performance, boasting an adjusted R^2 of 0.71 and a markedly lower AIC compared to all other models. This underscores MGWR as a more fitting local technique for elucidating COVID-19 variation in the region, given its capacity to consider spatial heterogeneity and local variability across space. Mansour et al. (2021) note that this study included confirmed COVID-19 cases data only and did not take into account suspected and asymptomatic cases, which might have influenced the analysis results but is equally an issue for all studies using COVID-19 rates including the present thesis.

Lin et al. (2020) investigated the relationship between socioeconomic determinants and the number of COVID-19 cases in 39 cities across China. The determinants included population, native population, population density, regional gross domestic product (GDP), per-person GDP, number of rural-to-urban migrants, proportion of rural-to-urban migrants, urbanization rate, proportion of tertiary industry, traffic capacity, per-person disposable income, number of hospitals, number of doctors, and number of travelers departing from Wuhan. The dependent variable was the number of COVID-19 cases. These socioeconomic determinants were chosen based on previous regional studies in the area. To assess the correlation between the number of COVID-19 cases and each independent determinant, scatter plots with regression lines were used to test for linearity. Then they used Spearman's correlation, which is robust in the presence of non-normality, to measure the strength of the relationship between the determinants. The Spearman correlation method was used because some data were not normally distributed. They selected independent determinants that had correlation coefficients that were $p < 0.15$ for inclusion within a multiple stepwise regression model. The determinants included in the final model were the number of travelers from Wuhan, population, native population, GDP, per capita GDP, number of hospitals, number of rural-to-urban migrants, traffic capacity, and per-person disposable income. However, a model based on the number of travelers from Wuhan and rural-to-urban migrants could explain 83% of the variance in the number of COVID-19 cases (adjusted. $R^2 = .833$). They did not test the model residuals for normality and so it is unclear if the model was mis specified. It is also important to note that using rates as the dependent variable, rather than counts, would have been more appropriate, as it would adjust for differences in population size. Additionally, they used regression curves to test the linearity assumption, which is not an appropriate test. Correlation analysis or biplots can provide more reliable results. Finally, increasing the pre-specified type one error rate to 0.1 to include more determinants in the regression analysis, indicates a potential methodological flaw, as this may result in identifying more significant relationships that are in fact due to chance, inflating the Type-I error rate, leading to erroneous conclusions.

Wang et al. (2021) investigated the temporal and spatial properties of COVID-19 in China and its driving factors. The dependent variable was the cumulative confirmed number of COVID-19 cases, whereas the determinants were population flow network data, air quality, precipitation, wind speed, temperature, school and workplace closures, public event cancellation, limiting gathering size, closing public transportation, stay-at-home requirements, government response strictness index, and national and international travel control (Wang et al., 2021). They found that the pattern of COVID-19 exhibited self-similarity globally using Moran's I and that some regions showed clustering of high values

and low values in addition to hot (high-low) and cold (low-high) cities using local Moran's I. Next, using SatScan (Kulldorf 1997; Kulldorf, n.d.), the daily cases were input into a Poisson regression model that identified periods through time at each specific location (where and when) that had high numbers of cases. To test the correlation between confirmed COVID-19 cases and each determinant, Wang et al. (2021) used Spearman's rank to measure the degree of correlation between two variables. The results of these analyses indicate that the cumulative number of confirmed COVID-19 cases possessed significant global spatial autocorrelation and a few clustered (High-High or Low-Low) regions across the study area, along with a few outliers (High-Low or Low-High) on different temporal scales. The natural determinants most highly correlated with the spread of COVID-19 cases in China were population migration (positive relation to COVID-19), air pollution concentration (positive), and temperature (negative relationship), while for the Government Response Strictness index factors, travel control had the biggest impact on disease spread. One issue is that this study established that the data was spatially autocorrelated yet proceeded to examining the bivariate correlations using Spearman's rank correlation, believing, in error, that a non-parametric statistic is not influenced by spatial dependency in the input variables. Therefore, spatial dependency in the two variables being compared can lead to inflated correlation coefficients because nearby or neighboring observations are more similar than what would be expected if they were randomly distributed. This inflation leads to the possibility of more Type I errors, namely rejecting a true null hypothesis of no relation. The author's could have used a measure that accounts for spatial correlation such as Lee's L (Lee 2004a).

Kuznetsov and Sadovskaya (2021) investigated the spatial variation of COVID-19 cases in Kazakhstan. The purpose of this study was to discover the geographical patterns of COVID-19 cases and highlight the spatial clusters of the disease. The data used for this research consisted of the confirmed number of COVID-19 cases and the population data in each administrative polygon unit (cities [n=3], oblasts [n=14] (akin to Canadian provinces), and rayons [n=174] (the smallest spatial unit of census). For each administrative unit, COVID-19 case counts were aggregated and polygon adjacency was determined using Queen's case, whereby any polygons touching the border of a given polygon are considered neighbours. They used the case locations themselves as pointsobjects and used the average nearest neighbor index to show that the cases were spatially clustered across all oblasts and cities. Unfortunately, it was obvious that the case locations were clustered within and near to cities, as seen in their first figure showing a map of cases and counts by oblast. They should have used a null hypothesis appropriate for clustered data, e.g., a spatial random process that generates clustered locations rather than testing a clearly clustered process against a completely spatial random process. In other words, the cases are clearly clustered and not random and testing against randomness is not required, whereas testing against a spatial point process model that generates clustered processes can provide information about the parameters of the point pattern distribution. Then they used global Moran's I to check the spatial dependencies of COVID-19 case counts within the polygon units. Kuznetsov and Sadovskaya (2021) have also used Getis-Ord Gi* and LISA analyses to test for spatial autocorrelation of confirmed COVID-19 counts between the study area's neighborhoods. Although they proposed that local population density could be a crucial factor influencing the epidemiological trend of the disease, their omission of rate computation during spatial autocorrelation statistics undermines the practical utility of their findings. The use of counts fails to account for the

variability in population size across administrative units, diminishing the relevance of the study's conclusions. This was overall a poorly executed study. This study was a univariate exploration of COVID-19 counts and while the authors suggest potential determinants, they did not produce any models.

A study in New York City detailed statistical insights into COVID-19 testing rates, positivity rates, and proportion of positive tests within at the Zip Codes level of geography (Cordes and Castro 2020). Using global Moran's I, their study revealed a strong positive spatial autocorrelation for all three outcomes, with Moran's I values of 0.698, 0.695, and 0.707 for testing rate, positivity rate, and proportion of positive tests, respectively. They also employed the SpatScan statistic (Kulldorff, M. 1997; Kulldorf, n.d.) to identify clusters with relative risks for each testing rate variable. They also used Pearson correlation to test for significant associations between the testing rate outcomes and various demographic and socioeconomic factors. Specifically, there were negative correlations found for COVID-19 rates and the Asian race, use of public transportation, education, non-citizens, and median income, while a positive correlation was found with rent (≥50 percent of income spent on rent). The proportion of positive tests showed strong positive and negative correlations with black and white populations, respectively. These findings highlight the disparities in COVID-19 impact and/or testing access across different demographic and socioeconomic groups in New York City. However, like Wang et al. (2021) they interpreted the Pearson correlations without correcting for spatial autocorrelation and, thus some of the relations found could be due to inflated Type I errors.

Castro et al. (2021) examined the dynamics of spatial dispersion of COVID-19 in Brazil in relation to socioeconomic indicators. The dependent variables were the COVID-19 incidence rates and mortality rates in each Brazilian municipality, however, only the incidence rates are considered here, as they are COVID-19 rates like the ones analyzed in this thesis. Their determinants were "illiteracy rate in people over 18 years old, Gini index, average income per capita, percentage of the population living in households with a density greater than two people per bedroom, proportion of the population in the household with a bathroom and running water, social vulnerability index (SVI), municipality human development index (MHDI), demographic density, and coverage of primary health care." (Castro et al. 2021). They stabilized their COVID-19 rate data using empirical Bayesian smoothing; a common technique used when small areas with large numbers of cases unduly effect regression parameters. The application of Moran's I indicated moderate positive and significant spatial autocorrelation for incidence rates ($I = 0.50$; $p < 0.05$). Local Moran's I found a few significant high-high and low-low clusters as well as high-low and low-high hot and cold spots respectively. Their multivariate regression model showed poor prediction accuracy with adjusted $R^2 = 0.20$. All variables showed a positive and statistically significant relation to COVID-19 rates except for the illiteracy rate of people over 18 years old and the proportion of the population at home with a bathroom and running water exhibited a negative impact on COVID-19 on incidence rates. Local Moran's I showed that the model residuals possessed significant spatial dependence and so the authors used Geographically weighted regression (GWR) for their final model achieving an adjusted $R^2 = 0.45$ and a lower AIC than the OLS method. The GWR beta coefficients have the same direction of relation as does the OLS model with the exception of the proportion of the population at home with a bathroom and running water, which becomes positive but close to zero. No significance test is shown for the GWR regression

model. In the GWR model, whose residuals were not tested for lack of spatial dependency, the strongest associations with incidence rates were with the SVI, MHDI, and proportion of the population with a density > 2 (Castro et al. 2021).

Han et al. (2021) analyzed the spatial distribution of COVID-19 incidence rates and their relationship with environmental determinants in Beijing. They first investigated the spatial clustering pattern of the COVID-19 using spatial autocorrelation analysis. They also utilized Spearman correlation analysis to characterize the relationship between COVID-19 cases and environmental determinants, and then applied GWR to examine the impact of these determinants on the spatial distribution of COVID-19 cases. The independent variable was COVID-19 cases, while the nine determinants selected that reflect the intensity of human activities and the likelihood of human contact in the region were population density, distance to Xinfadi market, distance to the hospital, distance to business sites, distance to educational facilities, distance to traffic facilities, distance to shopping sites, distance to parks, and distance to restaurants (Han et al. 2021). The use of global spatial autocorrelation was employed to assess the spatial dependence of COVID-19 cases in Beijing. LISA was then applied to identify local differences across the study area. The results of the Moran's I tests were evaluated for statistical significance using 999 Monte Carlo simulations. In addition, Spearman correlation was utilized to examine the strength of the linear relationship between determinants. Only significantly correlated determinants ($p < 0.05$), based on Spearman's rank correlation were used for GWR modeling. The results of this analysis suggest that population density and distance to the Xinfadi market were the most influential determinants in explaining the development of COVID-19 cases in Beijing (Han et al. 2021). In this examination, both the dependent variable and independent variables exhibited spatial autocorrelation. Consequently, employing Spearman's rank correlation coefficient to assess model inclusion for determinants would exaggerate the perceived significance, leading to the incorporation of variables that lack genuine linear relationships when accounting for spatial autocorrelation. Furthermore, the authors did not take into account multiple testing when using Local Moran's I but did derive local significance based on conditional Monte-Carlo simulations. In addition, there was no clear explanation of why certain determinants were included in the analysis.

A prominent theme across these studies is the identification and understanding of patterns of COVID-19 incidence and risk. This was achieved through mapping and statistical analysis, as demonstrated in the studies by Banu et al. (2012) and Siljander et al. (2022), which pinpoint peak transmission times and infection rates in specific locales. This approach is pivotal for targeted public health interventions and resource allocation that could potentially reduce the impact of any airborne disease, not just COVID-19 where, at least in the West is well under control.

The nature of the COVID-19 data varied from point to ecological level (polygon) in different studies. Additionally, some used incidence rates which are specific to a time-period and only count the number of new cases divided by the population at risk. Others use crude rates, referred to as rates, which are the number of infections over a time period including both new and past infections divided by the population at risk. One study (Kuznetsov and Sadovskaya 2021) erroneously used case counts, which would be permissible if the methods they used allowed count data to make sense in the context of the

analyses undertaken. In this research, crude rates are what this thesis uses for analysis because that is what we have available, and these are the sum of the number of all cases over the time period divided by the population at risk.

Methodologically, there's a consistent reliance on a common set of statistical models. Tools like Poisson regression, Ordinary Least Squares, Geographically Weighted Regression, and Spearman's correlation, despite the obvious issues, are commonly employed. These models are instrumental in deciphering how different determinants, such as socioeconomic and environmental determinants associate with COVID-19. By finding these determinants it becomes possible to profile census units that might be more vulnerable and target these for preparatory measures. The studies by Lin et al. (2020) and Castro et al. (2021), which explore how socioeconomic factors correlate with disease incidence, are good examples of this approach.

While most studies that used regression methods assessed overall relationship between determinants, some relied on GWR to automatically account for spatial heterogeneity due to spatial autocorrelation. Unfortunately, none of these studies tested the residuals of their regression models, be it OLS or GWR or Poisson regression and variants, for spatial autocorrelation. If spatial autocorrelation is present within modeled residuals from any regression model, then the regression coefficients are in fact biased, or so-called mis specified. In this current research, particular attention was given to producing unbiased models, testing residuals for spatial autocorrelation, and choosing models that eliminated the issue.

Determining common factors that are associated with disease rates suffer from data limitations and exclusions, as noted by Tang Vieira and Shearer (2022), can impede the comprehensiveness and accuracy of findings. The complex interplay of socio-demographic, environmental, and other factors also introduce layers of complexity, challenging the isolation of specific causal relationships. This complexity is compounded by the variability in methodologies across studies (e.g., see Table 5 for a summary) and varied geographic locations (China, Bangladesh, Brazil, and Finland). Cultural and place specific variables can hinder generalization of findings from one locale to another. As such, the generalizability of the studies reviewed here, and even more recent ones not reviewed suggest that localized studies might not accurately represent broader contexts and that models are not transferable from one region to another.

At the time of writing in 2022 there were few case studies using spatial methods to analyze COVID-19 patterns within the literature and thus, if this literature review was to be re-done today, a more comprehensive set of determinants and even methods could possibly be used to address the research questions. Research on COVID-19 patterns increase daily as researchers capitalize on the unique data that was produced during the pandemic, and this equates to a fast-moving field of research. In any fast-moving field what was written within a reasonable timeframe could seem out-of-date by the time it is published given the duration of a master's thesis.

In conclusion, unlike the studies described in this section, this research is primarily concerned with methodological research questions and compares both bottom-up and top-down approaches to understanding which health determinants are associated with COVID-19 rates in Ottawa. We first build a regression model top-down using determinants derived

using deductive logic based on the literature reviewed up to early 2022. Next, we compare that deductive approach to an inductive or bottom-up approach of data mining allows the identification of patterns and trends to uncover the strongest potential determinants of COVID-19 rates. The use of both approaches provides a thorough understanding of the importance of the underlying determinants in relation to COVID-19 in Ottawa, and by identifying these in this geographic region, it will enable the potential development of targeted interventions to mitigate the impact of future pandemics.

Chapter 2: Study Area, Data, and Methods.

1. Study Area

Ottawa, the capital city of Canada, is situated at the confluence of the Rideau River and Ottawa River on the eastern side of the province of Ontario, bordered by Gatineau, Quebec to the north, with the Ottawa River flowing between the two provinces. The city of Ottawa is divided into 111 neighborhoods by the Ottawa Neighbourhood Study (ONS) (M. P. Parenteau et al. 2008; Ottawa Neighborhood Study n.d.), which serve as the study area for this research on the evolution of spatial patterns of COVID-19 (Figure 1).

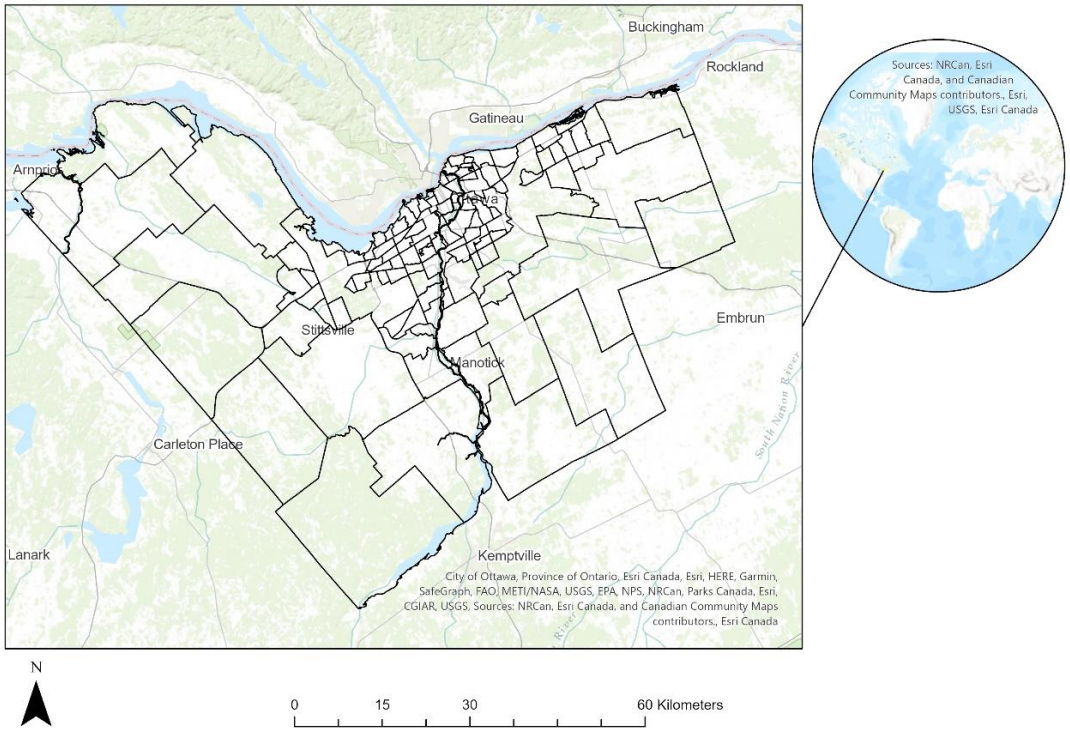


Figure 1. Location of the study area and neighbourhood geometries.

It is important to note at this point that other spatial units exist within which the analysis could have been done, including census units. However, the ONS boundaries are used by Ottawa Public Health (OPS) for reporting public COVID-19 rates. The only other spatial unit of public reporting is city Wards which are in the teen's and much larger. Moreover, the ONS boundaries were created for analyzing health outcomes with the goal of having within each neighbourhood a large enough population to support stable crude rate calculations (M. P. Parenteau et al. 2008). Due to the Modifiable Areal Unit Problem (MAUP) (Dark and Bram 2007; M.-P. Parenteau and Sawada 2011; A S Fotheringham and

Wong 1991) the analyses in this research are only applicable to the Neighbourhood scale and findings would differ at any other level of geographic scale.

2. Data and Exploratory Data Analysis (EDA)

All the data used were free and publicly available online from various sources. All analyses will be applied to the cumulative COVID-19 infection rates from July 2020 to December 2021.

Spatial Data: The files were produced using the Ottawa Neighborhood Study (ONS). The data were accessible online through the ONS website (Ottawa Neighborhood Study, 2023) or the Open Ottawa website (City of Ottawa, 2023). It contains the Ottawa neighborhood boundaries and population estimations for each neighborhood.

Of the 111 neighborhoods in Ottawa, nine were excluded from the analysis due to the absence of COVID-19 reporting or or small populations. These exclusions include Beechwood and Notre-Dame cemeteries, Carleton University, Greenbelt, Hunt Club South, Orleans industrial zone, Kinburn, LeBreton development, and Wateridge village (refer to Figure 3). The decision to exclude Beechwood and Notre-Dame cemeteries is grounded in the fact that these areas do not serve as residential spaces, being primarily cemeteries. Likewise, Carleton University was omitted from consideration. This decision stemmed from its status as a study and work institution, where a minimal census population of around 200 people could result in disproportionately high COVID rates if even a single case were reported. Consequently, this could create a scenario of elevated rates in a non-residential area solely because the registered population is limited to a few hundred people, while the effective non-census registered population is substantially larger due to the inclusion of students. However, the data we have is from a time when the daytime population was rather low due to lockdowns.

The exclusion of Greenbelt is based on its nature as a green space extending south of Ottawa's urban core, devoid of substantial residential populations. Regarding Hunt Club South (mostly airport), Orleans industrial zone (car dealerships), Kinburn (large rural area with only 23 cases), Leberton development (not completed and a population 519 persons in 2016), and Wateridge village (255 people in 2016) all contain low populations and so lead to unstable rates because of large variance. Moreover, these neighborhoods were omitted due to consistently having fewer than 5 reported COVID-19 cases monthly throughout the study timeline. Although the Ottawa Neighborhood Study neighbourhoods were created to include a minimum of 2000 persons per neighbourhood, in the above cases that was not possible. Thus, due to the small populations, even 5 cases can destabilize the analysis of cumulative rates compared to the average neighbourhood in Ottawa.

Health outcome data/Dependent variable:

COVID-19 monthly and cumulative rates were furnished by Ottawa Public Health (OPH) from the provincial Case and Contact Management (CCM) system (Ottawa Public Health, 2021). While the COVID-19 data contain monthly counts, particularly early in the

pandemic, most neighborhoods have zero cases monthly. There are numerous months between July 2020 and December 2021 during which there are fewer than 5 infected individuals in some neighbourhoods. This study will use only the cumulative rates of COVID-19, since data aggregated over time should mask any small-scale stochastic variation that might affect the generalizability of the results. This kind of data may be seen as mitigating the month-to-month fluctuations resulting from random variations in case detection efforts. However, it is important to note that public messaging on COVID-19 prevention measures evolved over the study period. These health interventions undoubtedly introduce bias to monthly incidence data but are expected to have a comparatively lesser impact on cumulative rates, given that the total sum incorporates all cases across all months.

The COVID-19 dataset analyzed in this study encompasses data from July 2020 to December 2021 for Ottawa neighborhoods, excluding cases linked to retirement homes and outbreaks in long-term care facilities (LTCFs). Notably, individuals living in long-term care homes were not excluded from the total population considered in calculating infection rates. This method introduces a potential bias towards lower effective infection rates in specific neighborhoods that contain LTCFs. Conversely, incorporating case counts from long-term care facilities within the respective neighborhoods would disproportionately elevate rates compared to those in the general population. By way of explanation, at the geographic level of Ottawa Neighbourhoods, the minimum population is over 2000 people with some well over 16,000 people, one of the criteria for their design of Ottawa Neighbourhoods was to have a minimum population in each neighbourhood that would increase rate stability for health geographic analyses (M. P. Parenteau et al. 2008). Consequently, the COVID-19 rates provided by Ottawa Public Health would be slightly conservative in neighbourhoods that contain long-term institutions. That is to say, the raw rates are the number of COVID-19 cases divided by the population. As such the population is effectively larger for neighbourhoods with long-term facilities leading to a bias therein towards smaller rates. That outcome is better than including the high case-numbers that occur due to direct contagion within the long-term care facilities, as including those would substantially affect neighbourhood rates and artificially increase the overall impact on community risk (Tang, Vieira, and Shearer, 2022). Finally, the data released by public health did not include long-term facilities and we have no information on those infections, since the rate data was meant to assess community risk and not LTCF risk. These aspects related to the removal of LTCF cases will be further discussed in the subsequent sections, to transparently address any implications for the interpretation of the infection rates.

Census Data: The socioeconomic determinants identified in the literature concern the economy, health, education, standard of living, environment, and general demographics. From the literature review, ten independent determinants (Table 1) that were identified in the reviewed literature as associated with COVID-19 rates were selected. Most of these determinants are from custom tabulations by Statistics Canada. A custom tabulation is produced using the same methods that are used for reporting the National Household Survey (NHS) within dissemination areas and census tracts but instead computed within the custom geography of the ONS boundaries.

Table 1. Description of the ten COVID-19 explanatory determinants, references to study using each definition, and data sources in Ottawa, Canada.

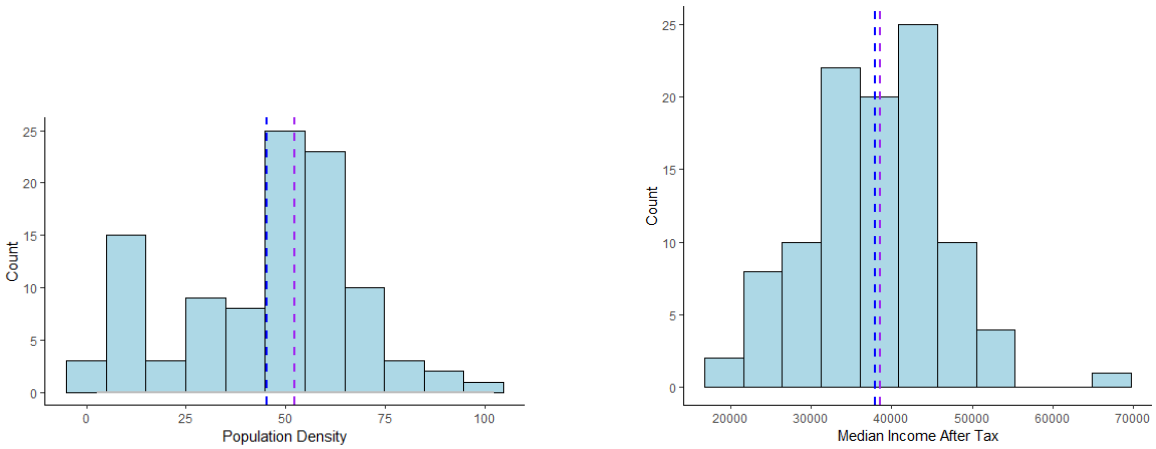
Determinants	Relevant Study	Definition	Source	Transformation
Population Density	(Mansour et al. 2021)	Number of persons per km2 in a neighbourhood.	Ottawa Neighborhood Study. 2016	Square root. Didn't reach normality but improved linearity.
Jobs Density	(Tang, Vieira, and Shearer 2022)	Total number of jobs in a neighbourhood per Km^2.	City of Ottawa: 2016 City of Ottawa Employment Survey	Logarithmic (base 10). Didn't reach normality but improved linearity.
Median income after tax	(Siljander et al. 2022)	Resident after-tax income refers to total income minus income taxes of the person during a specified reference period. Reported for population aged 15 years and over in private households.	Statistics Canada Census of Population; Ottawa Community Data Consortium, Community Data Program of the Canadian Council on Social Development. 2015	No need for transformation. Variable is already normally distributed.
Percentage of people with no high school diploma	(Siljander et al. 2022)	The category High school diploma or equivalency certificate includes persons who have completed the requirements for graduation from a secondary school or an equivalency certificate, but no postsecondary certificate, diploma, or degree. Reported for population aged 25 to 64 years in private households.	Statistics Canada Census of Population; Ottawa Community Data Consortium, Community Data Program of the Canadian Council on Social Development. 2016	Square root. Didn't reach normality but improved linearity.

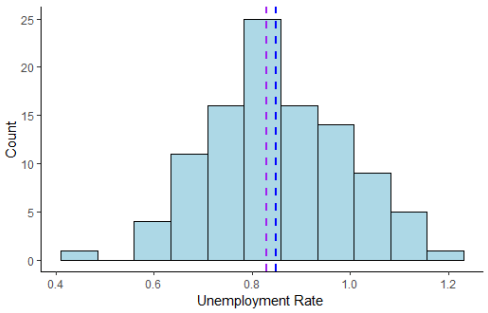
Unemployment rate	(Siljander et al. 2022)	The definition of "unemployed" refers to individuals who, during the week of Sunday, May 1 to Saturday, May 7, 2016, were not working for pay or were self-employed, were available to work, and either actively looked for work in the past four weeks, were on temporary lay-off with plans to return to their job or had definite plans to start a new job within four weeks. This definition applies to people aged 15 and over living in private households.	Statistics Canada Census of Population; Ottawa Community Data Consortium, Community Data Program of the Canadian Council on Social Development. 2016	Logarithmic (base 10). Reached normality.
Percentage of people with a bachelor level degree or above.	(Tang, Vieira, and Shearer 2022)	This category includes individuals aged 25 to 64 years who reside in private households and have earned one of the following: a bachelor's degree, a university certificate or diploma above the bachelor's level, a degree in medicine, dentistry, veterinary medicine, or optometry, a master’s degree, or a doctorate.	Statistics Canada Census of Population; Ottawa Community Data Consortium, Community Data Program of the Canadian Council on Social Development. 2016	Square root. Variable transformation did not achieve normality but transformation improves linearity.
Percentage of people who take public transit to work	(Tang, Vieira, and Shearer 2022) (Wang et al. 2021)	The main mode of commuting refers to the primary way in which individuals aged 15 years and over, who reside in private households and have worked at any time since January 1, 2015, travel between their homes and their workplaces.	Statistics Canada Census of Population; Ottawa Community Data Consortium, Community Data Program of the Canadian Council on Social Development. 2016	No transformation could improve its linearity.
Percentage of people over the age 65 years old.	(Mansour et al. 2021)	Age refers to the number of years a person has been alive up to their last birthday, or in relation to a specific reference date.	Statistics Canada Census of Population; Ottawa Community Data Consortium, Community Data Program of the Canadian Council on Social Development. 2016	Logarithmic (base 10). Reached normality.

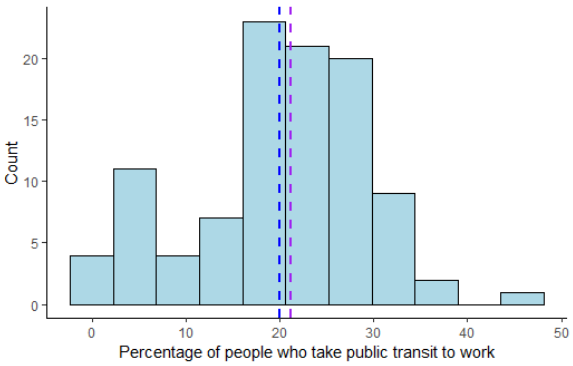
Percentage of newcomers (2011-2016)	(Tang, Vieira, and Shearer 2022)	Year of immigration refers to the year in which the immigrant first obtained landed immigrant or permanent resident status.	Statistics Canada Census of Population; Ottawa Community Data Consortium, Community Data Program of the Canadian Council on Social Development. 2016	Square root. Didn't reach normality but improved linearity.
Average household size	(Tang, Vieira, and Shearer 2022)	Household size refers to the total number of individuals living in a private household.	Statistics Canada Census of Population; Ottawa Community Data Consortium, Community Data Program of the Canadian Council on Social Development. 2016	Squared transformation. Reached normality.

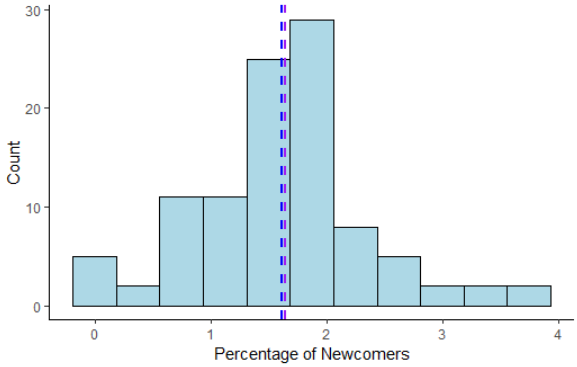
Note. The dependent variable is the cumulative COVID-19 rate in each Ottawa neighborhood per 100,000 population. A logarithmic transformation was applied to the dependent variable to create an approximate normal distribution, transformations were performed on the independent determinants to achieve less skewed distributions and to help achieve linear relationships.

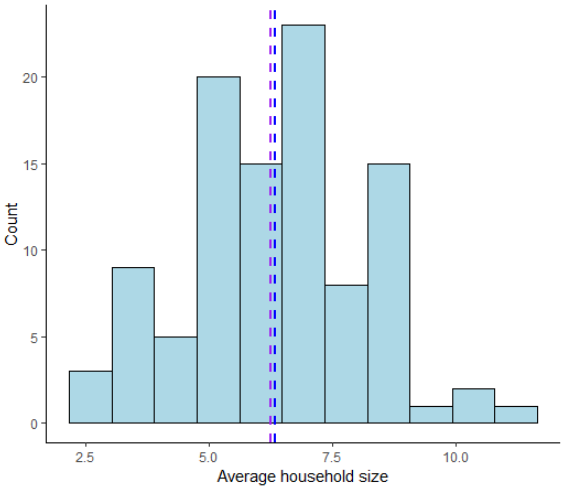
The dependent variable, cumulative COVID-19 rates, was logarithmically (base 10) transformed to achieve normality due to right-skewness. Transformation of the dependent variable to normality is a soft assumption for regression analysis, it is more likely that by transforming the variable to normality that the residuals will be more normal, which is a strong assumption of regression. Certain independent determinants were also transformed to create approximately normal distributions (Table 1, Figure 2). These transformations were performed to improve the linear relations, which can help the regression analysis meet assumptions on the residuals. Although a normal distribution is not required for the independent determinants in regression analysis, it can help achieve linearity between the cumulative rates and the determinants, and provide more robust results.

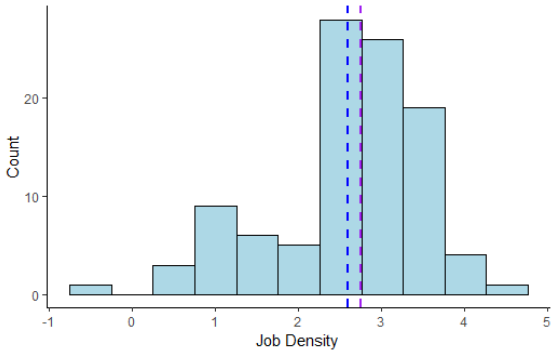


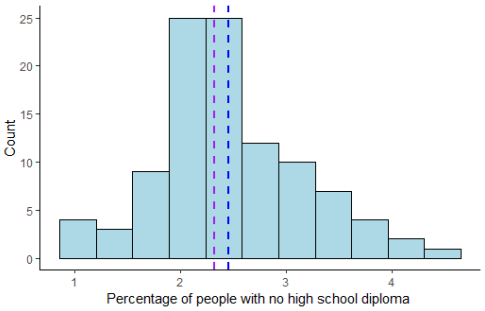

Count
Unemployment Rate


Count
Percentage of people who take public transit to work


Count
Percentage of Newcomers


Count
Average household size


Count
Job Density


Count
Percentage of people with no high school diploma

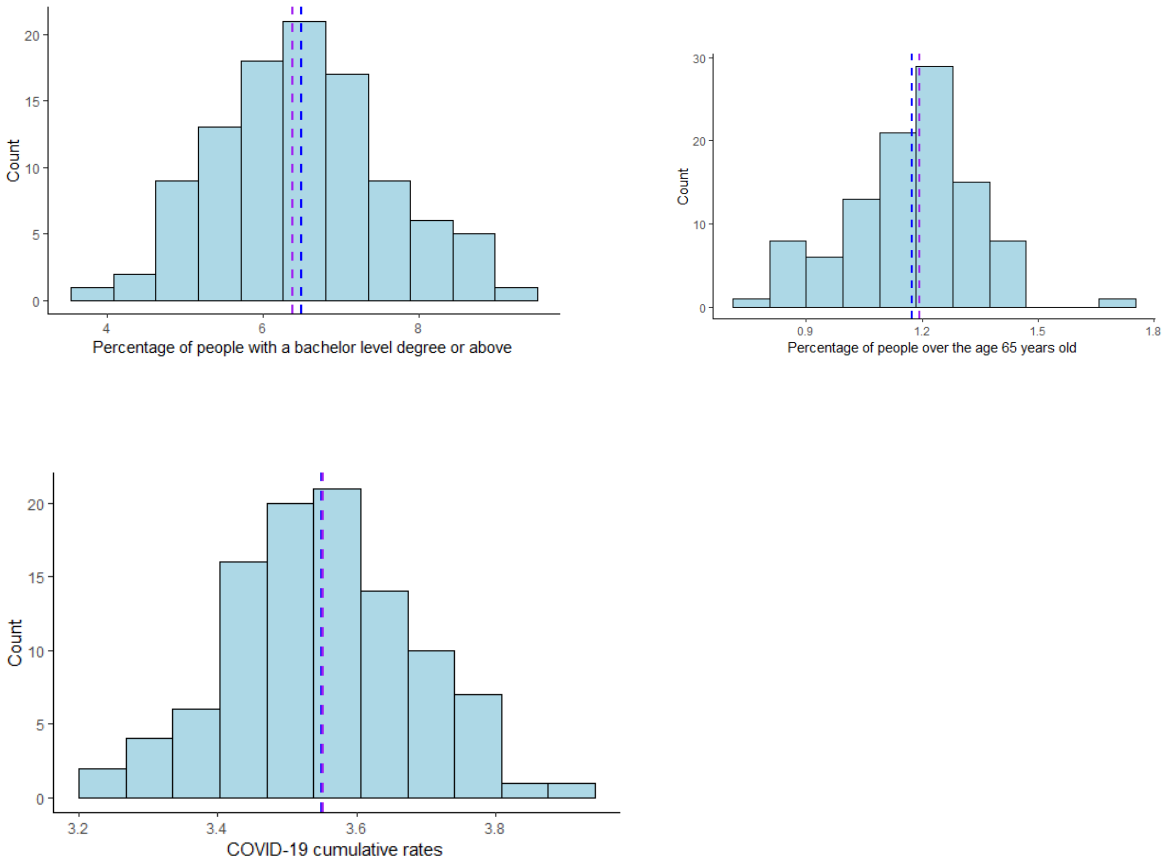


Figure 2. Histograms display the distribution of COVID-19 cumulative rates and relevant determinants. A mean dashed line in blue and a median dashed line in purple are included, the determinants are either transformed to approximate normality or non-transformed.

After applying a series of standard transformations on the determinants, normality was reassessed. One of the commonly used criteria for normality is the comparison of the mean and median of a dataset. If the mean is approximately equal to the median, this suggests that the data are symmetrically distributed around the average value like in in the Gaussian distribution. Skewness and kurtosis are also considered. A skewness of approximately zero indicates that the data does not have extreme values, whereas a kurtosis of approximately three suggests a bell-shaped distribution. The Shapiro-Wilk test (Shapiro and Wilk 1965) was used to test the normality of the determinants. The shapiro.test() function in R (R Core Team, 2023, version 4.2.2) returns a *p* and W statistic. If the *p* is greater than alpha = 0.05, the variable in question is approximately normally distributed because the null hypothesis that the variable is normal cannot be rejected. The final variables and their Shapiro-Wilk outcomes before and after transformation are illustrated in Table 2.

Table 2. Shapiro-Wilk test p-value and summary statistics for each variable. Summary statistics (mean, median, skewness, and kurtosis) were calculated after applying a transformation to each determinant. N/A means that transformations made the determinant more significantly different from a normal distribution and so no transform was applied.

Variable name	*p* before transformation	*p* after transformation	Kurtosis	Skewness	Mean	Median
Population Density	0.00	0.00	2.58	-0.33	45.14	52.21
Median Income After Tax	0.08	N/A	4.00	0.24	37957	38556
Unemployment Rate	0.00	0.51	3.05	0.00	0.84	0.82
Percentage of people who take public transit to work	0.01	N/A	2.87	-0.23	19.96	21.20
Percentage of Newcomers	0.00	0.00	3.94	0.28	1.61	1.64
Average household	0.03	0.24	2.52	-0.04	6.33	6.25
Job density	0.00	0.00	3.82	-0.98	2.59	2.74
Percentage of people with No high school diploma	0.00	0.02	3.13	0.51	2.46	2.32
Percentage of people with a Bachelor level degree or above	0.09	0.94	2.76	0.06	6.49	6.37
Percentage of people over the age 65 years old.	0.00	0.10	3.44	-0.1	1.17	1.19
COVID-19 cumulative rates	0.00	0.91	2.86	0.02	3.55	3.55

The median income after tax, unemployment rate, average household size, percentage of people with a bachelor's degree or above, percentage of people over the age of 65, and COVID-19 cumulative rates follow a normal distribution after transformation, as indicated by a $p > 0.05$, according to the Shapiro-Wilk test (Table 2). For determinants that did not achieve normality according to the Shapiro-Wilkes test, transformations were applied to create less skewed distributions, which should help achieve linearity with COVID-19 cumulative rates. The only exception being the percentage of people who take public transit to work, where each transformation created more significantly different distributions than the untransformed variable.

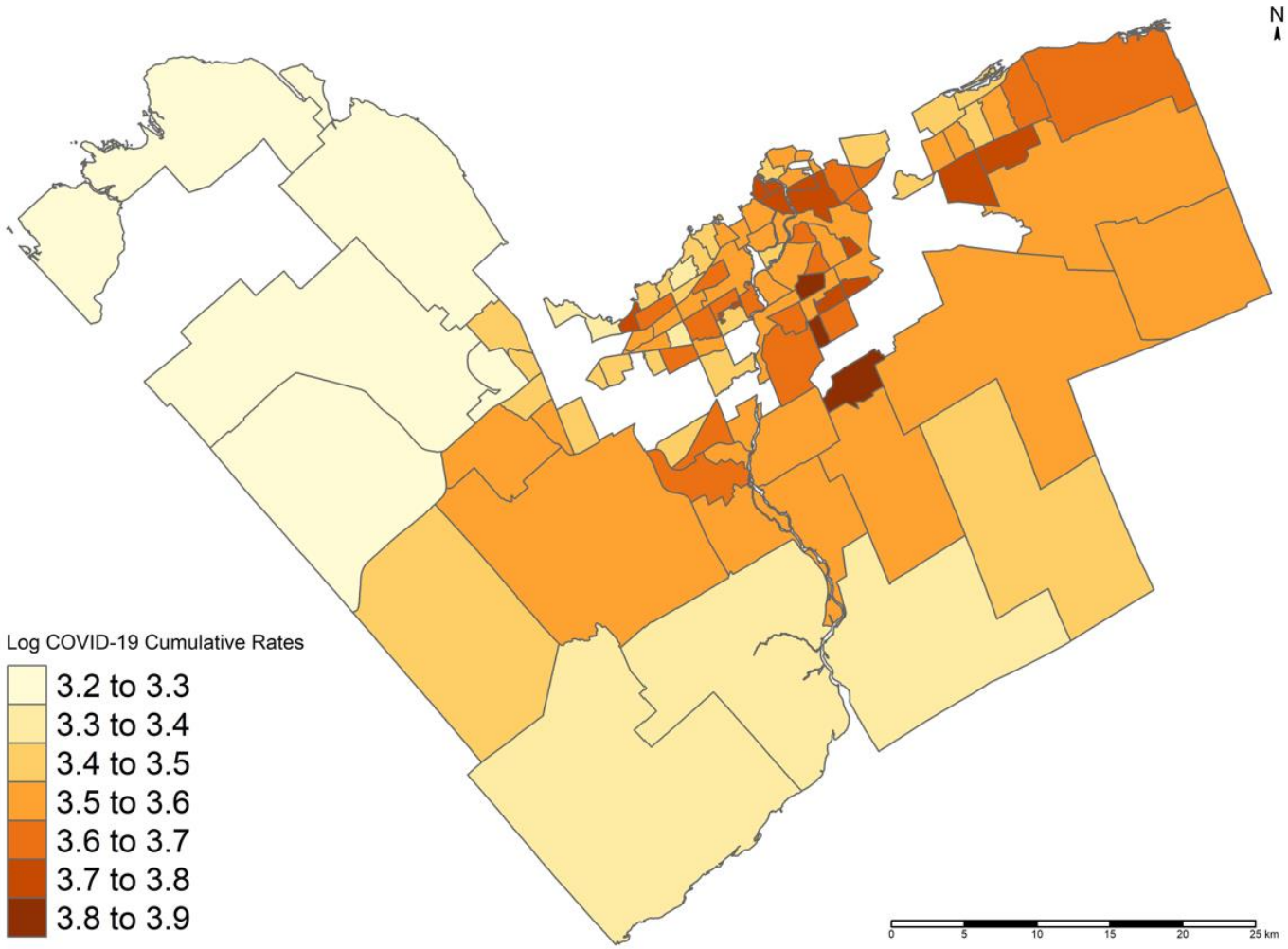


Figure 3. Distribution of COVID-19 cumulative rates in Ottawa neighborhoods with transformation applied. Whitespace within the boundaries indicate those neighbourhoods excluded from the analysis as explained in-text.

COVID-19 cumulative rates in Ottawa exhibit spatial variations across different neighborhoods. Most of the low COVID-19 rates are concentrated on the west and south sides of the city, exhibiting a pattern of lower COVID-19 incidence. Conversely, the highest COVID-19 rates are predominantly located on the east and north sides of the city. (Figure 3).

Figure 4. The strength of the linear relationships between the cumulative rates of log-transformed COVID-19 rates and the ten socioeconomic determinants, with the transformations applied per Table 1, the upper right diagonal shows the Pearson's R values between determinants, while the lower left diagonal displays the relationship between determinants in scatterplots with a linear trend in grey/black, the Pearson's correlation coefficient measures the strength of this relationship, in the upper diagonal with green indicating a positive linear relationship and pink indicating a negative linear relationship, the intensity of the color reflects the strength of the relationship, with darker shades representing a stronger relationship and lighter shades representing weaker relationships. See Table 3 for full determinant names based on abbreviations herein. For a larger version, refer to Appendix 1.

Table 3. Variable used in Figure 4.

Abbreviation in Figure 4	Full Determinant name
COVID_Rates	COVID-19 cumulative rates
Income	Median income after tax
Seniors	Percentage of people over 65 years old
Bachelors_or_above	Population of people with a bachelor's level degree or above
Avg#household	Average household size
Job_Density	Job Density
Pop_Density	Population density
no_HS_diploma	Population of people with no high school diploma
Unemployment_Rate	Unemployment rate
public_transit	Percentage of people who take public transit to work
Newcomers	Percentage of newcomers

Pearson's R correlation coefficient was used to explore the linearity of relations between the determinants and COVID-19 rates. This method does not consider the spatial autocorrelation that is present between the independent and dependent variables. It is used only to roughly assess the directions of variable relations and their general strength but not their significance. The percentage of newcomers has the strongest positive correlation with the COVID-19 cumulative rates in Ottawa (Pearson's r=0.57), followed by both the percentage of people who take public transit to work and the unemployment rate (r=0.52). On the other hand, the median income after tax has the strongest negative correlation (r=-0.53). The average household size and the percentage of people with a bachelor's degree or above show a weak negative correlation (Figure 4). Potentially significant collinearity was observed between certain determinants, including a negative relationship between the percentage of people with no high school diploma and median income after tax (r=-0.80) and a positive linear relationship between the percentage of people who take public transit to work and population density (r=0.79). Next, a negative relationship (r = -0.61) emerged between median income after tax and the percentage of people using public transit. In this case, where overall, higher income suggests lower public transit dependence. However, classic non-linearity is exhibited in this pattern which shows a scatter plot with an inverted-U shape. The relation shows in increasing dependence on public transit up to ~15% and after that the relationship starts to shift negatively.

To address the issue of multicollinearity among socioeconomic determinants, the variation inflation factor (VIF) was calculated based on an OLS model including all determinants. VIFs are usually pre-computed in order to identify potential issues with multicollinearity so that a set of non-collinear variables can be included in a model going forward.

Table 4. VIF values for socioeconomic determinants based on a single OLS regression, including all determinants for assessing VIF. The VIF will also be checked during actual model development.

Variable	VIF
Unemployment rate	3.44
Percentage of people with a Bachelor level degree or above	5.64
Percentage of people who take public transit to work	5.23
Average household	3.15
Job density	4.28
Median Income After Tax	6.23
Percentage of people with no high school diploma	5.62
Population density	3.95

Percentage of people over the age 65 years old.	2.02
Percentage of Newcomers	4.18

Assessing multicollinearity between the determinants in multiple regression models is crucial. A general rule of thumb is that VIFs exceeding ten may indicate serious multicollinearity, which requires further investigation (Menard 2002). However, in this case, the VIFs did not exceed ten, indicating that multicollinearity was not a major concern with this set of determinants (Table 4).

3. Methods

The methods selected to address the research questions for statistical modelling are derived from the literature review section. At the time of writing in early 2022, the available literature on spatial analyses of COVID-19 was small. A summary of methods and techniques at the time, the number of studies using said method and the methods themselves are summarized in Table 5.

Table 5. Different methods used in different studies sorted by the total number of papers using each method.

	OLS	Global Moran's I	Local Moran's I	Correlation	GWR	MGWR	Poisson regression	ANN	Getis-Ord Gi	GAM	SEM	SLM
Banu et al							X					
Siljander et al	X	X	X		X	X						
Tang et al										X		
Mansour et al	X					X					X	X
Lin et al	X			X								
Wang et al		X	X	X			X					
Kuznetsov and Sadovskaya		X	X					X	X			
Castro et al	X	X	X		X							
Han et al	X	X	X	X	X							
Sum	5	5	5	3	3	2	2	1	1	1	1	1

Similar to the spatial analysis study of COVID-19 and its socioeconomic factors in Finland (Siljander et al. 2022), this study addresses the first research question that regards understanding the pattern of COVID-19 by using global and local spatial autocorrelation (Anselin 2010). Also, Multiscale Geographically Weighted Regression (MGWR) maps (A. Stewart Fotheringham, Yang, and Kang 2017a) also provide local information on the determinants and COVID-19 by neighbourhood.

Global Moran's I is used to test if there is significant spatial autocorrelation of COVID-19 rates by providing an assessment of self-similarity or self-dissimilarity in the pattern of rates across Ottawa. Positive spatial autocorrelation, in which similar COVID-19 rates are close together in space, will be contrasted with negative spatial autocorrelation, in which dissimilar rates are close together in space. Spatial autocorrelation is absent if COVID-19 rates are randomly dispersed across Ottawa neighborhoods. The application of Moran's I assumes independence of COVID-19 rates among neighborhoods. Importantly, it is designed to detect spatial autocorrelation without presupposing a specific spatial clustering pattern. It also assumes a stationary mean (absence of trend), and that the variable being measured is normally distributed. Moran's I statistical significance is, therefore, generated through a Monte Carlo simulation that involves randomly rearranging COVID-19 rates across the study area and measuring the Moran's I value and repeating this shuffling of all COVID-19 values 9999 times. For each shuffle of COVID-19 rates, the outcome measure of *I* is stored to create a reference distribution under the null hypothesis of no significant spatial autocorrelation. Statistical significance of a spatial measure is then based on a one or two-tailed test of the observed value of Moran's I against the reference distribution. The global Moran's I value ranges from -1 to +1 and reflects the presence of spatial autocorrelation within the COVID-19 rates. A value of -1 indicates dissimilar rate clustering, whereas a value of +1 indicates similar rate clustering. A value near zero indicates the absence of spatial autocorrelation, and the magnitude of I reflects the strength of the spatial autocorrelation, if Moran's I is statistically significant (*p* smaller than the prespecified type 1 error rate, by convention, alpha = 0.05).

For all spatial analyses, a contiguity-based spatial weights matrix was defined between neighborhood polygons based on shared boundaries. This is called Queen's case contiguity and this matrix served as an input parameter for all spatial techniques used in R for analysis, including the moran.mc() function (Cliff and Ord, 1981) from the spdep package (R. Bivand 2022) (version 4.3.0), and all spatial regression models, including MGWR which uses a larger neighborhood that can differ between determinants. This is explained in a later section.

Since the global Moran's I does not provide specific information about which statistical units are similar or different from their neighbors, Local Indicators of Spatial Association (LISA) will be employed like local Moran's I (Anselin 2010). This research will use local Moran's I to explore the pattern of COVID-19 rates for clustering as well as spatial outliers that are either larger or smaller than their immediate neighbourhoods. When COVID-19 rates are similar in magnitude within the neighborhoods surrounding the one being tested, this results in positive local spatial autocorrelation, meaning high-high or low-low values. A neighborhood with a high (or low) value surrounded by other first-order neighbors with high (or low) values is considered a positive local spatial autocorrelation. Conversely, a neighborhood with a low value surrounded by high-valued first-order neighbors is known a

cold spot, and a neighborhood with a high value surrounded by first-order neighbors with lower values is called a hot spot, and both types of 'spots' indicate local negative spatial autocorrelation in both cases (Anselin 2010). Only significant values, however, are considered for interpretation.

Utilizing a local Moran's I function (Appendix 2), that includes a false discovery rate (FDR) (Benjamini and Hochberg 1995; Chen, Feng, and Yi 2017) correction for *p*, controls the proportion of false positives in multiple hypothesis testing reducing the Type I error rate. The local Moran's function is most likely run as a two-sided test that makes no assumptions on the direction of spatial autocorrelation for rates in a given neighborhood. To ensure that the expected proportion of false positives in multiple hypothesis testing remains below the desired alpha level (usually set to $p <= 0.05$), the FDR method adjusts the critical value for each test based on its rank among all locations. For each location, the null hypothesis specifies that there is no local spatial autocorrelation for the neighborhood in question. This adjustment is based on the estimated proportion of false positives, determined by the number of spatial units. Consequently, the adjusted critical value for each hypothesis depends on its rank among all the simultaneous hypotheses that individually possess a $p <= 0.05$ (Benjamini and Hochberg 1995; Chen, Feng, and Yi 2017). Thus, the FDR method can control the expected FDR level while still allowing for the identification of true positive results, and in that sense the overall test has more statistical power. The Local Moran function used the same spatial weights matrix used in the Global Moran calculations, as the study area has not changed, and 9999 permutations for the Monte Carlo simulations to test for statistically significant results.

Next, Ordinary Least Squares (OLS) was applied, utilizing a backward stepwise regression approach, to analyze the determinants listed in Table 1, hypothesized to influence COVID-19 rates as derived from the literature. OLS is a linear statistical method used to estimate the strength of the association between COVID-19 and its determinants (James et al. 2013). OLS coefficients minimize the sum of squared errors, representing the difference between the observed and the estimated values. Another study that utilized OLS regression in a COVID-19 rates was Mansour et al. (2021) in Oman, who included various sociodemographic determinants. The application of OLS regression requires certain assumptions to hold, including the assumption of linearity between COVID-19 rates and the independent determinants, the absence of significant collinearity, and the presence of homoscedasticity in the residuals with no spatial autocorrelation (Burton 2021). The lm() function (Chambers, 1992) and the step() function (Hastie, T. J. and Pregibon, D. 1992) from the stats package in R were used for this purpose (R Core Team, 2023, version 4.3.0).

Backwards stepwise regression, also called backward elimination, is a is a technique for constructing regression models where variables are iteratively removed from a comprehensive initial set. This process focuses on eliminating variables that do not enhance the model's predictive accuracy, thereby refining and streamlining the model for better performance. This technique initiates with an ordinary least squares regression model that encompasses all potential determinants and proceeds by eliminating the determinant with the largest *p*-value. Then a model is created using the *n*-1 remaining determinants. Subsequently, the least significant variable is removed again, and the process repeats until there is no longer any change in the goodness-of-fit measure called the Akaike Information Criterion (AIC), which was computed at each step. The model with the lowest

AIC is generally the most parsimonious and therefore is chosen as the final regression model. The resulting model is then considered complete, and the remaining set of determinants deemed statistically significant (Draper and Smith 1998). In this sense, the technique creates a more parsimonious model than the original model that contained all determinates. There are rules of thumb regarding which of the stepwise alternatives to choose as the best model and that is usually determined by whether the subsequent model's AIC has changed by at least 1 point. Eventually in the last few steps, the AIC changes very little and by the principle of parsimony one chooses the last step since it has the least determinants and so satisfies the principle of Occam's Razor that the simplest explanation is the best explanation.

OLS residuals are tested for spatial autocorrelation using the lm.morantest () function (Cliff and Ord, 1981) from the spdep package (R. Bivand 2022). This test is not the same as regular Moran's I and specifically is applied to OLS regression residuals and not residuals of other non-OLS models. In the latter case, residuals can be tested using Global Moran's I. If spatial autocorrelation is found in the residuals of the OLS model, then the next step is to produce what is called a full Spatial Durbin error model (SDEM) (J. LeSage and Pace 2009). The basis for this choice is to model how COVID-19 rates in one neighbourhood are spatially dependent on their neighbours. The Spatial Durbin error model is a spatial regression model that accounts for spatial dependence in the data using a combination of constituent models, including a spatial lag model (SLM) and spatial error model (SEM) and an OLS model (Manski 1993). Unlike global approaches that assume a uniform spatial dependence structure for the entire dataset, the Spatial Durbin Error Model considers the determinants within the immediate neighborhood of interest, as well as those in lagged neighborhoods and residual variation, in modelling COVID-19 rates.

The utilization of regression methods is an essential aspect of the analysis, as the models are based on a deductive logic / top-down approach, using determinants selected from the literature that are more likely to provide insights to the COVID-19 pattern in Ottawa. The Spatial Durbin Model (SDM) and the Spatial Durbin Error Model (SDEM) are nested within what is called the full Manski model (Manski 1993; J. LeSage and Pace 2009), and are examined as simpler spatial alternatives to correct the OLS coefficients. The Spatial regression models are specifically designed to incorporate spatial relationships and provide unbiased coefficients when dealing with spatially autocorrelated data (Figure 8). The main difference between SDM and SDEM lies theoretically in the theoretically differing assumed source(s) or role of spatial autocorrelation present in the OLS model residuals.

The SDM assumes residual spatial autocorrelation is caused from dependence within both the shared determinants in nearby neighbourhoods and resulting shared COVID-19 rates. SDM accounts for the indirect or spillover effects that are present when simultaneous changes in the COVID-19 rates in one neighborhood are related to COVID-19 rates in adjacent neighborhoods (for whatever reason, common living or working conditions or even direct contagion) and likewise with census determinants. These spillover effects capture the influence that neighboring areas have on each other due to spatial interaction processes across space (J. LeSage and Pace 2009). The question arises as to how likely it is that COVID-19 rates in each neighborhood are directly influenced by rates of, for example, educational attainment in the same neighborhood as well as adjacent neighborhoods?

Spatial autocorrelation is present in almost all geographic data because the arbitrary polygon units used to measure COVID-19, in this case neighborhoods, are quite often not coincident with the physical process that gives rise to the variance in the COVID-19 rates across space. In other words, the pseudo-spillover arises simply because the independent or dependent variables natural scale of variation occurs over a larger (or smaller) region than a given neighborhood polygon can capture. In such cases, the SDEM model may be preferable because it considers spatial dependence in the residuals arising from either some unknown process like the non-concordance of process-scale interactions, or due to missing variables that were not included in the model itself (J. P. LeSage 2008).

Geographically Weighted Regression (GWR) is a method for analyzing spatial data and creates separate regression equations for each neighborhood in Ottawa. GWR accounts for the fact that the relationships between the dependent and independent determinants may vary from one spatial unit to another, rather than assuming that these relationships are uniform across the entire study area (Brunsdon, Fotheringham, and Charlton 1998).

To address this spatial heterogeneity, GWR creates a separate local beta parameter for each spatial unit derived from the values of COVID-19 rates regressed against the determinants within that neighborhood and its contiguous neighborhoods (Brunsdon, Fotheringham, and Charlton 1998). GWR fits a regression model that enables the individual spatial unit's coefficients to vary based on geographic location (Brunsdon, Fotheringham, and Charlton 1998). GWR is used to gain an understanding of the strength and direction of the relationships between determinants in different neighborhoods, since coefficients can be mapped to show local spatial variation. GWR has several assumptions, including the assumption of linearity, the residuals (errors) of the model are normally distributed, the variance of the residuals is constant across different values of the independent determinants (homoscedasticity), and that there is no spatial autocorrelation in the residuals (Brunsdon, Fotheringham, and Charlton 1998; Wheeler 2014; A. Stewart Fotheringham, Yang, and Kang 2017b). By employing GWR in the examination of the strength of the direction of the relationships between the determinants and COVID-19 rates, the objective is to uncover the spatial disparities and fluctuations that vary regionally in pattern (Brunsdon, Fotheringham, and Charlton 1998).

The estimation of the optimal bandwidth parameter is a crucial step in GWR. The bandwidth determines the operational spatial scale at which relationships between the determinants and COVID-19 rates are modelled (Gollini et al. 2015a; Rossiter 2022; A. Stewart Fotheringham, Yang, and Kang 2017b). The bandwidth parameter can be adjusted to create either a more sensitive model to local variations using a smaller bandwidth or a more global model using a larger bandwidth. Several other parameters must be set in this method, including the approach, kernel function, and adaptive options. The approach parameter can be set to either AIC or cross-validation (CV). In the CV approach, the optimal bandwidth value is then selected by minimizing the cross-validation error, allowing it to handle both global and local spatial structures (Wheeler 2014). The gwr() function from the spgwr package (version 0.6-35) (R. Bivand and Yu 2023) was utilized to construct a GWR model. Initially, cross-validation was employed using a fixed bandwidth and an Exponential kernel.

On the other hand, the AIC approach (Li et al. 2020) uses this goodness of fit to determine the optimal bandwidth value by finding bandwidth that produces the lowest AIC when testing different bandwidths and kernels. The optimal bandwidth value is selected by minimizing the AICc value (Li et al. 2020). There is no correction or consideration here for multiple testing; no statistical significance is available to determine if the drop in AIC from one distance to the next is due to chance or real effects. Next, the kernel specification can be specified as Gaussian, Exponential, Bisquare, Tricube, or Boxcar (Li et al. 2020; Bidanset et al. 2017) (see Figure 10 for examples of the kernel functional forms). A kernel is a function that generally operates by reducing the weights assigned to determinants as distance from the kernel center increases. In short, they operate by assigning weights to neighboring polygons based on their proximity. Because In MGWR, the kernel plays a critical role: It determines how weights are assigned to determinants used to predict COVID-19 rates. Second, different kernel functions will yield different spatial scales at which relationships between rates and determinants are optimal (Rossiter 2022; A. Stewart Fotheringham, Yang, and Kang 2017b; Bidanset et al. 2017).

The Gaussian kernel assigns weights based on the Euclidean distance between observations and the center of the kernel bandwidth (the effective mean of the Gaussian curve) (Rossiter 2022). The weights decrease as the distance from the center of the polygon of interest increases following the Gaussian shape, reducing the influence of observations further from the center on parameter estimates. However, the weights assigned to features that are distant from the center of the kernel have a negligible effect on the overall regression because, depending on the kernel, it either approaches zero weight or is zero if the kernel uses a cutoff distance/adjacency.

In contrast, the Bisquare kernel assigns a weight of zero to observations located outside the bandwidth. This eliminates their impact on the local regression estimate, reducing the effect of outliers while preserving sensitivity to local patterns (Rossiter 2022; Bidanset et al. 2017).

The Exponential kernel is more resilient to outliers, as outlier observations are assigned lower weights, reducing their influence on the regression estimates when compared to the Gaussian kernel (Rossiter 2022; Bidanset et al. 2017). The Tricube kernel, like the Bisquare kernel, is robust to outliers, but it has a wider bandwidth and assigns smaller weights to observations at the edge of the bandwidth (Rossiter 2022). This kernel is more likely to assign zero weights to these observations, resulting in a sharper decline in the weights assigned to observations compared to the Bisquare kernel.

Finally, the Boxcar kernel is the simplest kernel function, but its usefulness is limited as it does not consider the distance between observations (Rossiter 2022). The Boxcar kernel assigns a weight of 1 to all observations within the bandwidth and a weight of 0 to all observations outside the bandwidth. This approach limits the utility of the Boxcar kernel and it is rarely used in practice (Bidanset et al. 2017).

Furthermore, the bandwidth parameter can be specified as either a fixed or adaptive kernel. The adaptive kernel approach adjusts the size of the bandwidth based on the number of neighboring observations at which AIC is minimized (Li et al. 2020; Bidanset et al. 2017), whereas the fixed kernel sets a constant bandwidth size for all observations. In particular, the adaptive kernel method selects the number of neighboring observations to include in

the regression analysis and modifies the bandwidth size based on the density of the observations.(Lombard, Stern, and Clarke 2016). On the other hand, the fixed kernel method assigns a fixed distance or bandwidth size to all observations, regardless of the density of the data points (Terrell and Scott 1992).

One issue with GWR is that the bandwidth is optimized across all determinants and so the spatial scale of the relationship between COVID-19 and each determinant becomes a constant. MGWR provides a more realistic representation of the spatial scale of each determinant by allowing bandwidth to vary by determinant and so addresses that limitation within GWR (Li et al. 2020). In this research the gwr.multiscale() function (Yang, W. ,2014) was used from GWmodel package (Gollini et al. 2015b) to experiment with MGWR. To assess the effectiveness of the MGWR models cross-validation was implemented across all models (Wheeler 2014).

In the process of selecting the appropriate MGWR model, the first step was to ensure that the model is not biased or mis-specified. Although a model with a high adjusted R^2 value and the lowest AIC value may seem appealing, it is important to verify whether the MGWR assumptions are met. Spatial autocorrelation is often inflating these values when assumptions are violated. Two tests were conducted to verify the model's assumptions. Firstly, the normality of the residuals was checked using the Shapiro-Wilk test (Shapiro and Wilk, 1965) using the shapiro.test() function in R (R Core Team, 2023, version 4.2.2). Secondly, the Moran's I statistic was computed using the moran.mc() function (Cliff and Ord, 1981) from the spdep package (R. Bivand 2022) (version 4.3.0) to determine if the residuals exhibit any significant positive spatial autocorrelation.

Next, in addressing the last research question of whether a bottom-up approach could produce an effective model to predict COVID-19 rates, a Random Forest (RF) model was chosen because it is suitable for both regression and classification problems. RF is an ensemble learning technique that combines multiple decision trees (hence many decision trees make a forest) (Breiman 2001a). The popularity of Random Forest in data mining is due to its ability to handle large datasets, nonlinearities, missing values, and provide robust predictions even in the presence of irrelevant or noisy data. To develop the predictive model for COVID-19 rates using RF, over 338 demographic, environmental, and socioeconomic determinants were obtained from the same sources as the census data listed in Table 1. Random Forest splits the entire dataset into a training set and a test set, with the training set used to train the algorithm and the test set used to evaluate the model performance (ESRI 2023b; Suthaharan 2016; ESRI 2023a).

A decision tree is a widely used machine learning algorithm that helps in solving classification and regression problems. It operates by recursively splitting the dataset into subsets based on the most important attributes, eventually leading to the prediction of a target variable (ESRI 2023b; 2023b; Breiman 2001a). The process of creating multiple decision trees involves selecting the optimal point (condition) to divide each determinant in each tree separately in the training set using metrics usually referred to as information gain, such as Gini impurity or entropy. (Suthaharan 2016)

Once the model is trained, it can be used to predict the COVID-19 rates based on new data input socioeconomic determinants. In addition to its predictive capability, the algorithm

can identify the most important socioeconomic determinants that have the greatest importance in predicting COVID-19 rates.

The importance parameter in the randomForest() function, set to true, facilitates the calculation of variable importance measures. Specifically, the importance() function with type = 1 is employed to calculate the 'Mean Decrease in Accuracy'. (Liaw and Wiener 2002) sometimes called permutation accuracy importance (Strobl et al. 2007). The calculation involves permuting the determinants at each node from the out of bag (OOB) determinants (the data not used in constructing that tree) and then measuring how much the overall accuracy of the model decreases (Hoare 2018). A greater drop in prediction accuracy when that variable is withheld, determines the relative importance of that determinant in constructing the overall model performance or accuracy (Hoare 2018). The decrease in accuracy is then averaged across all trees to get the final importance score (Hoare 2018). The proximity argument is used to generate a proximity matrix that measures the similarity between observations based on their classification patterns in the random forest model (Liaw and Wiener 2002). This matrix reflects how frequently two observations are classified together in the same terminal nodes across all trees, enabling tasks such as cluster analysis and outlier detection (Liaw and Wiener 2002). The random forest model used in this research was trained using the training dataset, specifying 18 trees with importance and proximity calculations enabled. Additionally, a value of 83 was selected for the optimal number of determinants randomly sampled at each split.

Incorrect utilization of Random Forest models can result in poor model generalization (ESRI 2023b; 2023a; Bergstra and Bengio 2012). It is vital to thoroughly tune the model and choose the appropriate hyperparameters for each specific application to ensure reliable results. Several algorithms have been developed to simplify this process. The Grid Search algorithm is a simple and straightforward solution that attempts different combinations of hyperparameters and selects the best combination based on the model performance, usually as determined by the R^2 score (Bergstra and Bengio 2012).

The Random Search Algorithm operates similarly to the Grid Search, but instead of trying out all combinations, it randomly selects combinations for testing (Bergstra and Bengio 2012). Bayesian Optimization uses Bayesian principles to model the relationship between hyperparameters and the model performance and then suggests the next set of hyperparameters to test (Snoek, Larochelle, & Adams, 2012). These hyperparameters include the number of trees in the forest, the maximum depth of trees controls the number of levels in each decision tree, the number of features considered at each split refers to the number of socioeconomic determinants to be considered at each split of the decision tree branch, the minimum samples required to split sets the minimum number of determinants required in a node for a split to occur and move on to the next node, and bootstrap sampling creates multiple subsets of the original dataset. And for each subset, a decision tree is trained to reduce overfitting.

When trying to predict COVID-19 rates in Ottawa using a Random Forest model, it is crucial to address residual spatial autocorrelation in model residuals if it exists if so then the model parameters would be mis-specified (R. S. Bivand, Pebesma, and Gómez-Rubio 2013). Because the COVID rate data are spatially dependent, as are the determinants, analyzing such data frequently results in spatial autocorrelation issues in model residuals

akin to OLS, as previously described. A crucial aspect of this process is to guarantee that the residuals do not display any spatial dependency patterns, which can be visualized using a spatial plot or an error distribution plot. (R. S. Bivand, Pebesma, and Gómez-Rubio 2013). Furthermore, in a recent study, Hu et al. (2022) have utilized spatial eigenvectors generated through a contiguity-based spatial weights matrix to tackle spatial autocorrelation that determinants in a standard Random Forest model may not capture. By incorporating these spatial eigenvectors into their methodology, they have successfully mitigated a substantial portion of the spatial autocorrelation inherent in the data, leading to enhanced prediction accuracy. (Hu, Chun, and Griffith 2022)

In order to compare model outputs, difference maps are used whereby the observed COVID-19 rates have subtracted from the model predictions. The resultant difference is a map of the residuals, and we explore the degree of spatial structure therein for an individual model. Next, we compare prediction maps to observed COVID-19 rates and each model predictions to each other models' predictions using both Pearson's R and Lee's L measure of bivariate spatial autocorrelation (Lee 2001; 2004b), Lee's L is designed to assess the spatial association between two different variables. Lee's L integrates concepts from both Pearson's R, which measures linear correlation, and Moran's I, which assesses spatial autocorrelation (Lee 2001) and takes into account both spatial and in-situ correlations, offering a more nuanced understanding of how variables are related in a spatial context. In that sense, it controls for the inflation of Pearson's R when spatial autocorrelation is present.

This research uses a variety of spatial methods to investigate the pattern of COVID-19 and its relationships to health determinants. For example, to address the first research question regarding the spatial dependency of the COVID-19 and to identify spatial clusters and outliers, Global and Local Moran's I are employed. To address the second research question regarding the development and effectiveness of a statistical top-down modelling of COVID-19 based on the determinants identified in the literature, OLS and spatial regression models like the SEM are used, in addition to GWR and MGWR. To address the last research question, a bottom-up data-mining approach uses a machine learning Random Forest model to determine the socioeconomic determinants that are most important to predicting the cumulative infection rates of COVID-19 in Ottawa. All unbiased models were compared to determine which was most effective at predicting COVID-19 cumulative rates. OLS, GWR, and Random Forest analyses also aid in understanding the relationships between COVID-19 rates and the most influential socioeconomic determinants. Finally, a model intercomparison was completed to understand the spatial differences predicted by each model.

Chapter 3: Analysis and Results

Henceforward, the term 'COVID-19' is used to refer to the log-transformed values of the COVID-19 cumulative rates.

1. Spatial Dependence Analysis in the Pattern of COVID-19.

Global Moran's I test assesses the presence of spatial autocorrelation in the cumulative COVID-19 rates in Ottawa. The focus was on a one-tailed positive alternative hypothesis. This is because all previous studies found global positive spatial autocorrelation and that prior knowledge suggests that positive spatial autocorrelation of COVID-19 rates should also be found in Ottawa, as is often the case with almost all spatial data. Therefore, the alternative parameter of the moran.mc() function was set to "greater" to specifically search for evidence of positive spatial autocorrelation. The test yielded a statistic value of 0.39 and a p= 0.0001, indicating statistically significant moderate positive spatial autocorrelation in the cumulative COVID-19 rates in Ottawa. This test suggests that the rates of the disease are not occurring randomly by neighborhood and adjacent neighborhoods have similar values. Neighborhoods with high COVID-19 rates tend to be close to other neighborhoods with high rates, and neighborhoods with low rates tend to be close to other neighborhoods with low COVID-19 rates.

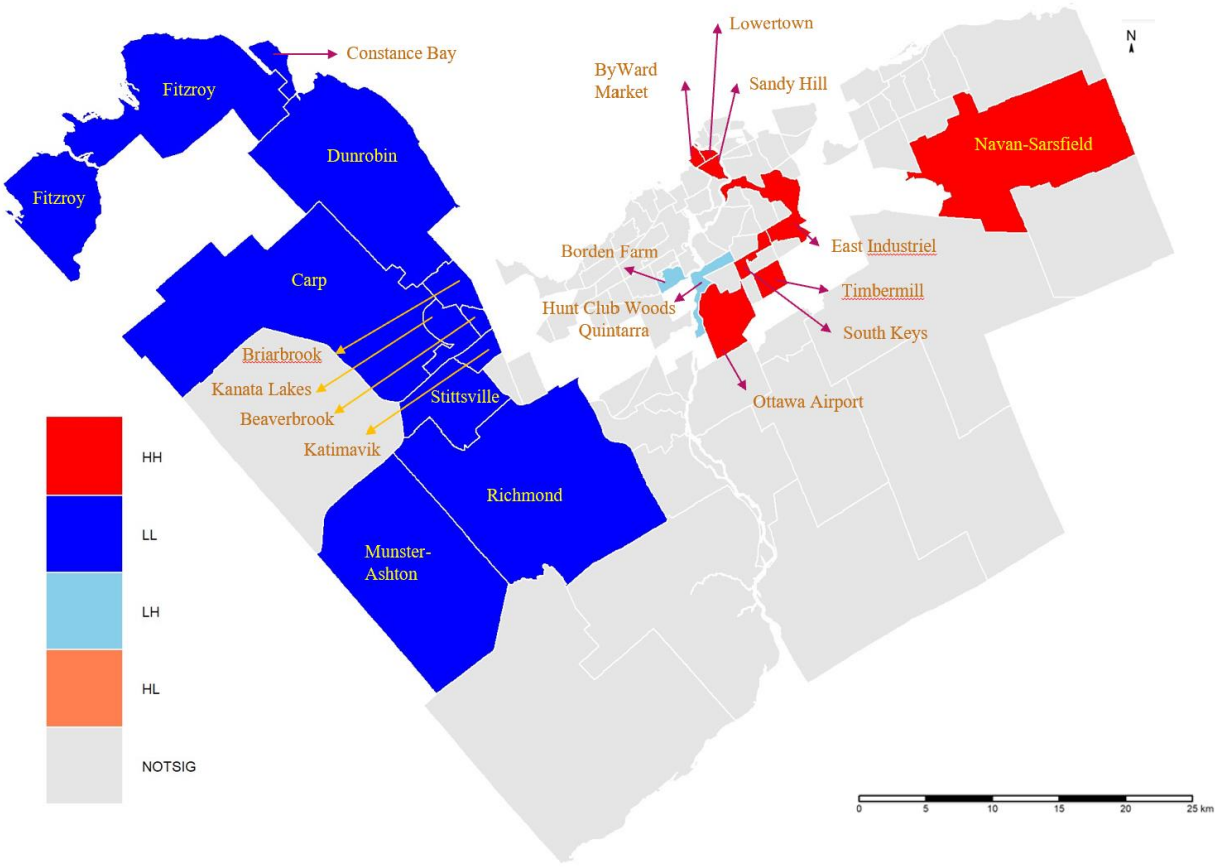


Figure 5. Local Moran's I results indicating types of clusters. These results used the FDR adjusted p for COVID-19 cumulative rates in Ottawa.

Next, because global Moran's I test does not provide which specific neighborhoods have similar or different COVID-19 rates from their neighbors, Local Moran's I is employed to identify which spatial units show statistically significant spatial autocorrelation.

The analysis of the spatial distribution of COVID-19 rates in Ottawa reveals notable patterns (Figure 5). Two clusters of positive local spatial autocorrelation are evident, one on the western side with low COVID-19 rates (blue), and another on the eastern side with high COVID-19 rates (red). Additionally, two cold spots have been identified in the central northern part of the city (light blue) whereas the remaining neighborhoods exhibit no statistically significant spatial autocorrelation (grey). The analysis revealed that some neighborhoods in Ottawa exhibited positive spatial autocorrelation in low COVID-19 cumulative rates, while others showed positive spatial autocorrelation in high COVID-19 rates. Specifically, Fitzroy, Constance Bay, Dunrobin, Carp, Briarbrook, Beaverbrook, Kanata Lakes, Katimavik, Stittsville, Munster – Ashton, and Richmond were found to contribute significantly to the (Low-Low) positive spatial autocorrelation. Conversely, Navan – Sarsfield, Byward Market, Lowertown, Sandy Hill, East Industrial, South Keys, Ottawa Airport, Timbermill exhibited positive spatial autocorrelation in the high COVID-

19 rates. Furthermore, two cold spots were: Hunt Club Woods – Quintarra, and Borden Farm.

2. The top-down approach to modeling the pattern of COVID-19: Multivariate Regression Analysis

Initially, all 10 determinants (see Table 2 for transformations) were included in the model, and stepwise regression was used to remove one variable at a time based on the AIC value of each model. At each iteration, the AIC value of the reduced model was computed, and the process was repeated until the model with the lowest AIC value was found.

Table 6. Stepwise Regression Results for Selecting Determinants of COVID-19 Rates

Step	Determinants Removed	Model Formula	AIC
1	NA, all determinants included.	COVID 19 Rates ~ Population Density + Median Income After Tax+ Unemployment Rate + Percentage people that take public transit to work + Percentage of Newcomers + Average household + Job Density + Percentage of People with No High School Diploma + Percentage of people with bachelor's degree and Above + Percentage of People Over 65 years old.	-467.66
2	Median Income after Tax	COVID 19 Rates ~ Population Density +Unemployment Rate + Percentage people that take public transit to work + Percentage of Newcomers + Average household + Job Density + Percentage of People with No High School Diploma + Percentage of people with bachelor's degree and Above + Percentage of People Over 65 years old.	-469.64
3	Unemployment Rate	COVID 19 Rates ~ Population Density + Percentage people that take public transit to work + Percentage of Newcomers + Average household + Job Density + Percentage of People with No High School Diploma + Percentage of people with bachelor's degree and Above + Percentage of People Over 65 years old.	-471.53
4	Population Density	COVID 19 Rates ~ Percentage people that take public transit to work + Percentage of Newcomers + Average household + Job Density + Percentage of People with No High School Diploma + Percentage of people with bachelor's degree and Above + Percentage of People Over 65 years old.	-473.4
5	Job Density	COVID 19 Rates ~ Percentage people that take public transit to work + Percentage of Newcomers + Average household + Percentage of People with No High School Diploma + Percentage of people with bachelor's degree and Above + Percentage of People Over 65 years old.	-474.92
6	Average household	COVID 19 Rates ~ Percentage people that take public transit to work + Percentage of Newcomers + Percentage of People with No High School Diploma + Percentage of people with bachelor's degree and Above + Percentage of People Over 65 years old.	-476.48

7	Percentage of Newcomers	COVID 19 Rates ~ Percentage people that take public transit to work + Percentage of People with No High School Diploma + Percentage of people with bachelor's degree and Above + Percentage of People Over 65 years old.	-477.51

The model with the lowest AIC value (-477.51) obtained through stepwise regression included four determinants: percentage of people who take public transit to work, percentage of people with no high school diploma, percentage of people over 65 years old, and percentage of people with a Bachelor level degree or above (Table 6).

Next, the direction and statistical significance of each determinant's relationship with COVID-19 was tested and the determinant's VIF value, to determine the impact of multicollinearity on the variance of the regression coefficients (Mansfield and Helms 1982) (Table 7).

Table 7. Multiple Linear Regression Results for COVID-19 Cumulative Rates and the determinants resulted from Stepwise Regression.

Variable	Coefficient	VIF	*p*
(Intercept)	3.35	-	<0.001
Percentage of people who take public transit to work	0.002	1.91	0.036
Percentage of people with no high school diploma	0.13	3.16	<0.001
Percentage of people with a Bachelor level degree or above	0.04	2.71	0.007
Percentage of people over the age 65 years old	-0.36	1.05	<0.001
AIC	-477.51		
R^2 / R^2 adjusted	0.551 / 0.533		
Shapiro-Wilk Test *p*	0.25		

All retained determinants are statistically significant and had a positive relation to COVID-19 rates. The only exception to this was the percentage of people over 65 years of age, meaning that as age increased, COVID-19 rates decreased. The adjusted R-squared explained approximately 53.3% of the variation in COVID-19 cumulative rates (Table 7). Based on the VIF values, there is no evidence of significant multicollinearity among the determinants, suggesting that all determinants are reliable in the regression model.

The normality and spatial dependency of the model residuals were also assessed. The histogram of the residuals was examined, and the Shapiro-Wilk test was used to test for normality (Shapiro and Wilk, 1965), resulting in a p= 0.25, indicating that the observed distribution is not significantly different from a Gaussian distribution.

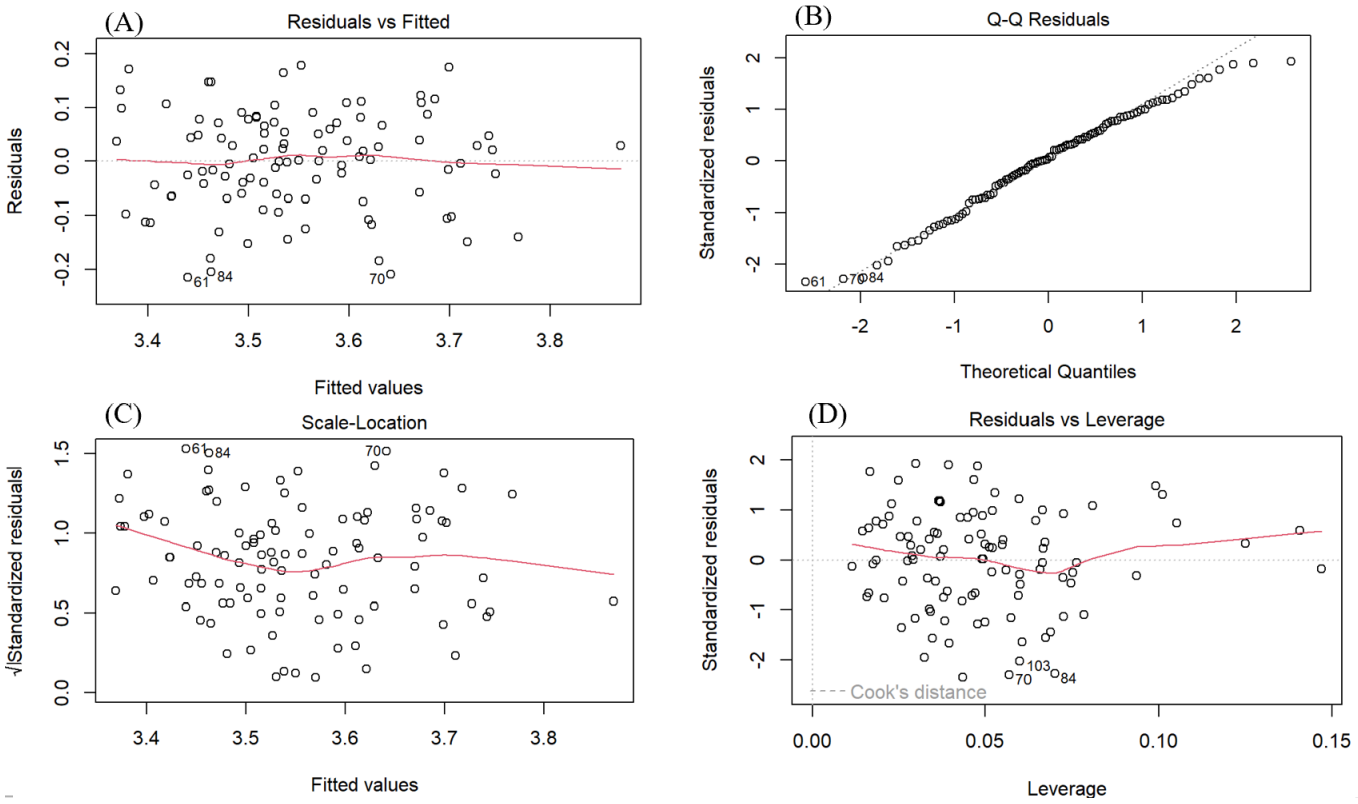


Figure 6. Residual Diagnostic Plots for Model Assessment: Residuals (A), Normality (B), Homoscedasticity (C), and Influential Observations (D), point 61 corresponds to Kanata Lakes, 70 corresponds to Britannia Village, 84 corresponds to Constance Bay, and 103 corresponds to Briarbrook.

Next, four residual diagnostic tests were conducted to examine residuals (Figure 6). The Residuals vs. Fitted plot (Figure 6a), ideally should indicate that the residuals exhibit a random scatter around y=0. A residual-fitted plot would illustrate a band of points along the x-axis at y=0 with a spread that is equal in the vertical on each side of y=0. In this instance, the residuals are randomly scattered above and below the horizontal zero with little bias positive or negative, thus indicating a good model fit. Additionally, three specific neighborhoods with small numbers are displayed on the graph (Figure 6a). The QQ residuals plot (Figure 6b) assesses normality by examining whether the points align roughly along a diagonal straight line, indicating a normal distribution. Although most of the residuals conform to the straight-line pattern, a few minor deviations are present, including the same neighborhoods observed in the Residuals vs Fitted plot. The scale-

location plot (Figure 6c) is used to check for homoscedasticity. The standardized residuals are not forming a clear pattern. Rather, they seem scattered around a horizontal line suggesting that the assumption of homoscedasticity is met and there is a linear relationship between the determinants and COVID-19 rates. Kanata Lakes, Britannia Village, and Constance Bay are also highlighted in this plot. The Residuals vs. Leverage plot (Figure 6d) is used to identify influential observations in the model. Britannia Village, Constance Bay, and Briarbrook (Figure 7) are flagged as high leverage points and may bias the regression coefficient even though they do not exceed Cook's distance of 0.5, they warrant a closer look.

The residual diagnostic plots indicate that the model assumptions, such as normality, homoscedasticity, and influential observations, are generally satisfied. However, some minor deviations occur in, and these require further investigation to determine if they are biasing the model parameters.

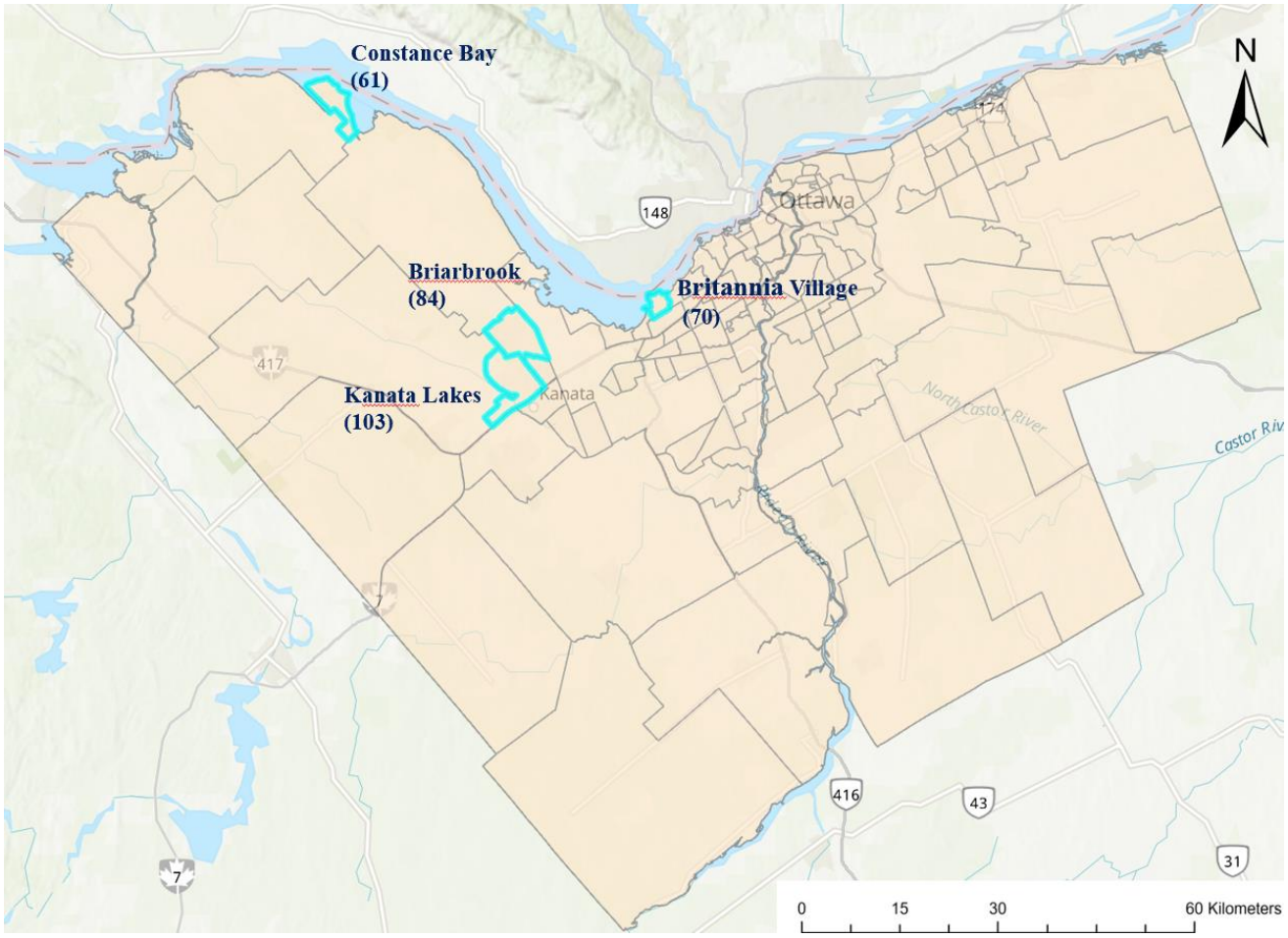


Figure 7. Residual Analysis: Map of Flagged Neighborhoods.

Kanata Lakes, Britannia Village, Constance Bay, and Briarbrook have the potential to be outliers, which may require exclusion from the analysis if they are found to be caused by human errors such as data entry or measurement errors (Figure 7). These neighborhoods demonstrate an unusually large relationship between the predicted and observed COVID-19. The Residuals vs Fitted panel (Figure 6a) illustrates this difference as negative, indicating that the predicted COVID-19 rate value is larger than the actual value, or that the prediction overestimates the actual value. Therefore, one or many of the determinants that contributed to the predicted value may be biasing the regression parameters.

Table 8. Neighborhood Comparison Table: Flagged Neighborhoods Values vs. average of The Entire Dataset.

	Variable Name										
Neighborhood	Pop Density	Income	Unemployment	Public Transit	Newcomers	Avg.household	Job Density	No HS Diploma	Bachelors and above	Over 65	COVID Rates
Kanata Lakes	1374.9	46643	6.1	16.5	4.3	2.8	452.1	1.7	67.8	16.9	1677
(Value – average)	(-1165.7)	(+8686)	(-1.3)	(-3.4)	(+1.2)	(+0.3)	(-1128.3)	(-4.9)	(+24.5)	(+1.0)	(-2048)
Britannia Village	3017.3	29761	9.5	34	4.9	1.8	880.3	12.7	37.4	24	2705
(Value–average)	(+476.6)	(-8196)	(+2.1)	(+14.1)	(+1.8)	(-0.7)	(-700.1)	(+6.1)	(-5.9)	(+8.1)	(-1020)
Constance Bay	377.1	36212	8.5	3	0	2.4	12.7	7.3	19.3	13.8	1811
(Value–average)	(-2163.6)	(-1745)	(+1.1)	(-16.9)	(-3.1)	(-0.1)	(-1567.7)	(+0.7)	(-24.0)	(-2.1)	(-1914)
Briarbrook	1671.7	41803	6.6	17.8	3.9	3.2	2721.9	4.0	53.6	7.19	2790
(Value–average)	(-868.9)	(+3846)	(-0.8)	(-2.2)	(+0.8)	(+0.7)	(+1141.5)	(-2.6)	(+10.3)	(-8.7)	(-935)
Entire dataset average	2540.7	37957	7.4	20.0	3.1	2.5	1580.4	6.6	43.3	15.9	3725

Note. The table shows the original values of each variable without any transformations at the top, followed by the difference between the neighborhood value and the average of the entire dataset. The use of (+) and (-) indicates whether the difference is above or below the average, respectively, highlighting the magnitude of the difference and facilitating the comparison.

Constance Bay exhibits the largest deviations from the dataset average, particularly in population density, percentage of people who take public transit to work, percentage of newcomers, job density, and cumulative rates (Table 8). In the west end of Ottawa, Kanata Lakes and Briarbrook, are two adjacent neighborhoods that exhibit no substantial deviation

from the values' average of the determinants of the complete dataset, except for job density. Despite this minor difference, their removal from the analysis lacks justification, and therefore, these points were retained. Briarbrook is associated with a large negative standardized residual, which indicates an overestimation of COVID-19 rates. However, its inclusion is justified since none of its determinants' average values significantly deviate from the average of the entire dataset. Therefore, it is not considered a high leverage outlier and will remain in the analysis as well. Moreover, the relatively lower COVID-19 rates in these neighborhoods compared to the entire dataset, with Constance Bay having the lowest cumulative rate, suggests that these neighborhoods may have unique combinations of determinants that influence the magnitude of COVID-19 rates.

Another OLS assumption is the lack of residual spatial autocorrelation. The local Moran for regression residuals yielded a Moran's I statistic of 0.34 ($p = 1.278\times10^{-7}$) for the OLS model. Due to the presence of weak, yet statistically significant positive spatial autocorrelation in the residuals of the regression model, the assumption of residual independence (e.g., no spatial autocorrelation in the OLS model residuals) was violated. As a result, the model parameters were biased or mis-specified and should not be used for further analysis. Therefore, spatial regression models were used to produce regression parameters and an unbiased model that considers spatial autocorrelation within the OLS residuals. The source of residual spatial autocorrelation is mostly due to the presence of spatial autocorrelation in one or more of the COVID-19 rates (already established) and the independent determinants.

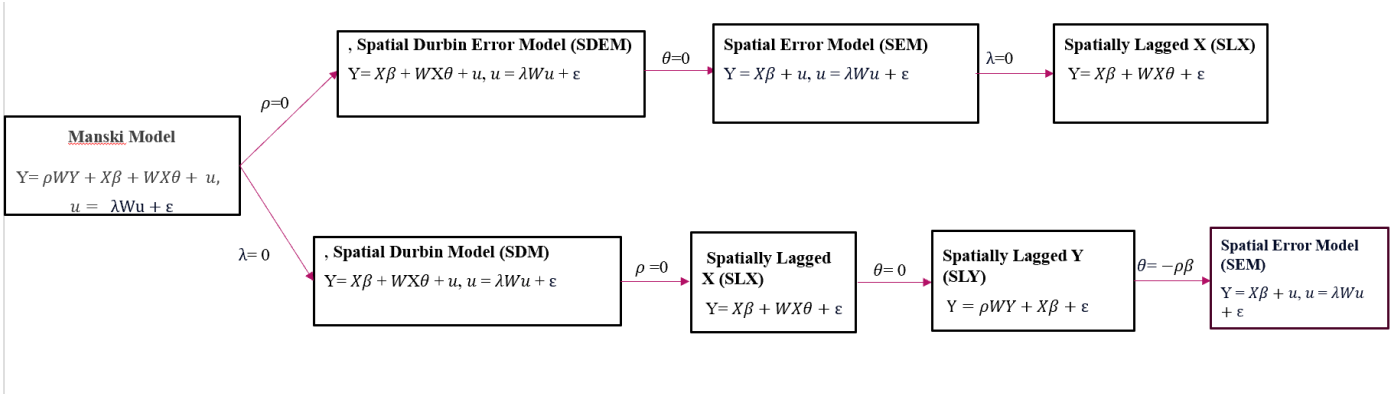


Figure 8. Nested Spatial Econometric Models: Manski, SDM, SDEM, SLX, SLY, and SEM. Adapted from (Burkey, 2018). Y: COVID-19 rates, W: spatial weights matrix, u: total error term, X: determinants, ρ: (Rho) spatial autoregressive parameter, ε: stochastic error term (noise), β: (Slope) coefficient associated with X, θ: (Theta) spatial lag effect coefficient, λ: (Lambda) spatial lag effect for the error term u.

In the analysis of COVID-19 rates, there are subjective interpretations regarding the mechanisms that lead to the spatial dynamics of its spread. Some may argue that the effect of the virus in one neighborhood can potentially ripple across the entire study area due to positive feedback mechanisms. Positive feedback occurs when COVID-19 rates in a given neighborhood cause (because of shared disease pathways via determinant values) lead to higher (or lower) rates in the neighborhood in question every time they change. This type of feedback loop is characterized by a self-reinforcing cycle, where an initial change leads to further change in the same direction in the neighborhood context. As such, high-rates of

COVID-19 in one neighborhood affect its first-order neighbors to move to high-rates, and so on. The higher rates increasing in the neighboring neighborhoods lead to even higher rates in the original neighborhood via amplification. When feedback is a reasonable assumption, then an SDM model would be appropriate. In Ottawa, the mechanisms that control this feedback are likely because of shared determinants that lead to the actual contagion that creates the rates, e.g, perhaps people in nearby neighborhoods work in the same places, visit the same places, and catch the virus.

For instance, this effect is also influenced by scale effects inherent in the process and determinants. Artificial boundaries between neighborhoods contribute to this phenomenon as they may not entirely encapsulate the factors contributing to high rates. In that case the starting model should be SDEM. While both perspectives are valid, there remains the dilemma of determining which model to start with. Therefore, both models were assessed, and their performance was compared to gain insights into the spatial dynamics of COVID-19 and census determinants on the OLS model.

To compare the SDM and the SDEM, both model's AIC values were compared. The AIC for SDM was found to be -31.29, while the AIC for SDEM was -29.85. Again, these are nested within the Manski model and so AICs are directly comparable. The model with the lowest AIC was chosen, which in this case was the SDM. Nested within the SDM are SEM and SLY, both models are simpler and contain the same parameters. The LR.Sarlm (Krämer & Sonnberger, 1986) function from the spatialreg package (R. S. Bivand, Pebesma, and Gómez-Rubio 2013; R. Bivand 2022; R. Bivand and Lewin-Koh 2016; R. Bivand and Piras 2015; Pebesma and Bivand 2023) was computed and found that there were no statistically significant differences between the nested SEM or SLY and the full SDM (respectively p= 0.2572, p= 0.8483). Therefore, for the sake of parsimony (Both SLY and SEM have fewer parameters than SDM) and the ease of SEM coefficient interpretation, SEM was preferred over SLY.

Without considering model parsimony, which favors simpler models over more complex ones when choosing among models that possess similar or near equal ability to explain the observed data. The Lagrange Multiplier test was conducted using the lm.LMtests function (Anselin, 1988) from the spatialreg package (R. S. Bivand, Pebesma, and Gómez-Rubio 2013; R. Bivand 2022; R. Bivand and Lewin-Koh 2016; R. Bivand and Piras 2015; Pebesma and Bivand 2023) (version 1.2-6) . However, both models SEM and SLY showed statistically significant results (respectively $p = 3.655\times10^{-6}$, $p = 1.656\times10^{-7}$) indicating that the robust forms of the Lagrange multiplier statistics need to be considered. Subsequently, for the robust forms, the Lagrange multiplier test yielded SEM with p= 0.5577 and SLY with p= 0.01205. Consequently, the results of the Lagrange multiplier diagnostic indicate that SLY is the more suitable model for the spatial dependence in the OLS model. That is consistent with the SLY's lower p when comparing SEM vs SLY to the SDM.

Given these findings, the SLY model would be the better quantitative choice over the SEM. However, given the fact that several unknown variables are likely omitted from the OLS model (the chosen set is based on recent studies in other countries/cities), there is sufficient reason to believe that there are unobserved determinants that are spatially correlated that directly impact the observed COVID-19 rates in Ottawa. For example, if rates are high in one neighborhood, they would also be high in adjacent ones, since

neighborhoods are artificial boundaries and do not coincide with the spatial scale of the socioeconomic processes associated with disease rates. Moreover, SEM has the fewest number of parameters and so is more parsimonious. In conclusion, SEM would yield an equivalent overall interpretation as the SLY (Table 9). However, the beta coefficients in SEM are directly interpretable, akin to OLS coefficients. On the contrary, SLY coefficients, being divided into direct and indirect effects that cannot be directly summed, pose challenges for straightforward interpretation.

There is no reason to believe that the observed spatial autocorrelation in COVID-19 rates between neighborhoods was due to in-situ spatial process of contagion within the neighborhood itself. Rather it is more likely that scale effects of similar socioeconomic determinants lead to similar healthscapes (Rainham et al. 2010) for people living in adjacent neighborhoods. The term "healthscape" denotes the combined physical and social environment shaping the health of individuals and communities. This encompassing concept includes elements such as healthcare facilities, public spaces, community resources, and social determinants of health like housing, education, and economic factors. These factors, encountered in everyday activities within the broader environment, play a pivotal role in determining COVID-19 infections (Rainham et al. 2010). From a healthscape perspective, people in similar socioeconomic circumstances may come into contact with more individuals in their daily healthscapes, elevating the risk of catching COVID-19 if people with similar socioeconomic status tend to work in service industries for example. Subsequently, the elevated risk under similar socioeconomic circumstances leads to neighborhoods with similar rates of the disease.

Table 9. Comparison of OLS and Spatial Model Results. Columns names correspond to models shown in Figure 8.

Dependent variable: COVID-19 Cumulative Rates						
	OLS	SEM	SLY	SLX	SDM	SDEM
% people public take public transit to work	0.002969*	0.002754*	0.00246*	0.001897	0.00219470	0.00228707
%people with no high school diploma	0.126786*	0.105978*	0.09893*	0.098632*	0.092503*	0.10465*
%people with bachelor level degree or above	0.039288*	0.032019*	0.0216630	0.019523	0.01959897	0.02723772
% people over 65 years old	-0.36253*	-0.34589*	-0.3665*	-0.40969*	-0.37973*	-0.3699*
Lagged % people public take public transit to work	NA	NA	NA	0.002332	-0.00053588	-0.0005397
Lagged % people with no high school diploma	NA	NA	NA	0.061597	0.00918042	0.06704737
Lagged %people with bachelor level degree or above	NA	NA	NA	0.015452	0.00463872	0.02340245
Lagged % people over 65 years old	NA	NA	NA	-0.101064	0.08692039	-0.0678941
Intercept	3.348299*	3.422488*	2.01433*	3.443418*	1.815261*	3.26911*
Lambda	NA	0.49608	NA	NA	NA	0.4698

Rho	NA	NA	0.43083	NA	0.46013	NA
Log likelihood	100.1	109.0631	111.0288	102.96	111.7169	110.9972
Adjusted R^2	0.53	0.63	0.64	0.54	0.65	0.64
Multiple R^2	0.55	0.65	0.66	0.58	0.67	0.65
Residual Standard Error	0.09398 on 97 DF	0.082 on 96 DF	0.081 on 96 DF	0.09259 on 93 DF	0.082 on 92 DF	0.083 on 92 DF
F-statistic	29.82 on 4 and 97 DF***	NA	NA	16.23 on 8 and 93 DF	NA	NA
LR test Value	-17.92*	20.078*	24.01*	NA	18.046*	16.606*
Wald statistic	NA	27.822*	27.664*	NA	22.464*	23.46*
$Sigma^2$	NA	0.0063985	0.0062834	0.0086	0.0061458	0.0062143
AIC	-477.51	-204.13	-208.06	-185.38	-201.43	-199.99

Note: * indicates $p \leq 0.05$.

Despite having the lowest AIC value, indicating its initial suitability, the OLS model was deemed inappropriate for the analysis due to violations of its underlying assumptions. Due to existing spatial autocorrelation in the OLS model, beta parameters are inflated. Consequently, the SLY model emerged as the most suitable approach, with the SEM model next (Table 9).

Focusing solely on the SEM model, all determinants have statistically significant coefficients, with positive relationships with COVID-19 observed for the percentage of people taking public transit, the percentage of people with no high school diploma, and the percentage of people with a bachelor's degree or higher. For instance, an increase of one unit in the percentage of people taking public transit to work is associated with an increase of 0.27% of the logged COVID-19 cumulative rates per 100,000 people. In contrast, for each one-unit increase in the percentage of people over 65 years old, COVID-19 cumulative rates are expected to decrease by approximately 29.24% (Table 9).

The spatial autoregressive parameter, Lambda (λ) determines the strength of spatial dependence among the observations. A value of λ close to 1 indicates strong spatial dependence, while a value close to 0 indicates weak or no spatial dependence. In this case, the value of λ is 0.49608, indicating moderate spatial dependence in the data. Using the LR.Sarlm function to compare the fit of data between SEM and its nested OLS model, the LR test value yielded a -17.92 with 5 degrees of freedom, and a $p = 3.04 \times 10^{-3}$. This indicates that the SEM model, which has a higher log-likelihood value of 109.0631 compared to the OLS model's log-likelihood value of 100.10, provides a significantly better fit to the data (Table 9).

To assess the model's performance, three maps were generated showing the predicted rates, the actual rates and the residuals or differences between the observed and predicted (Figure 9)

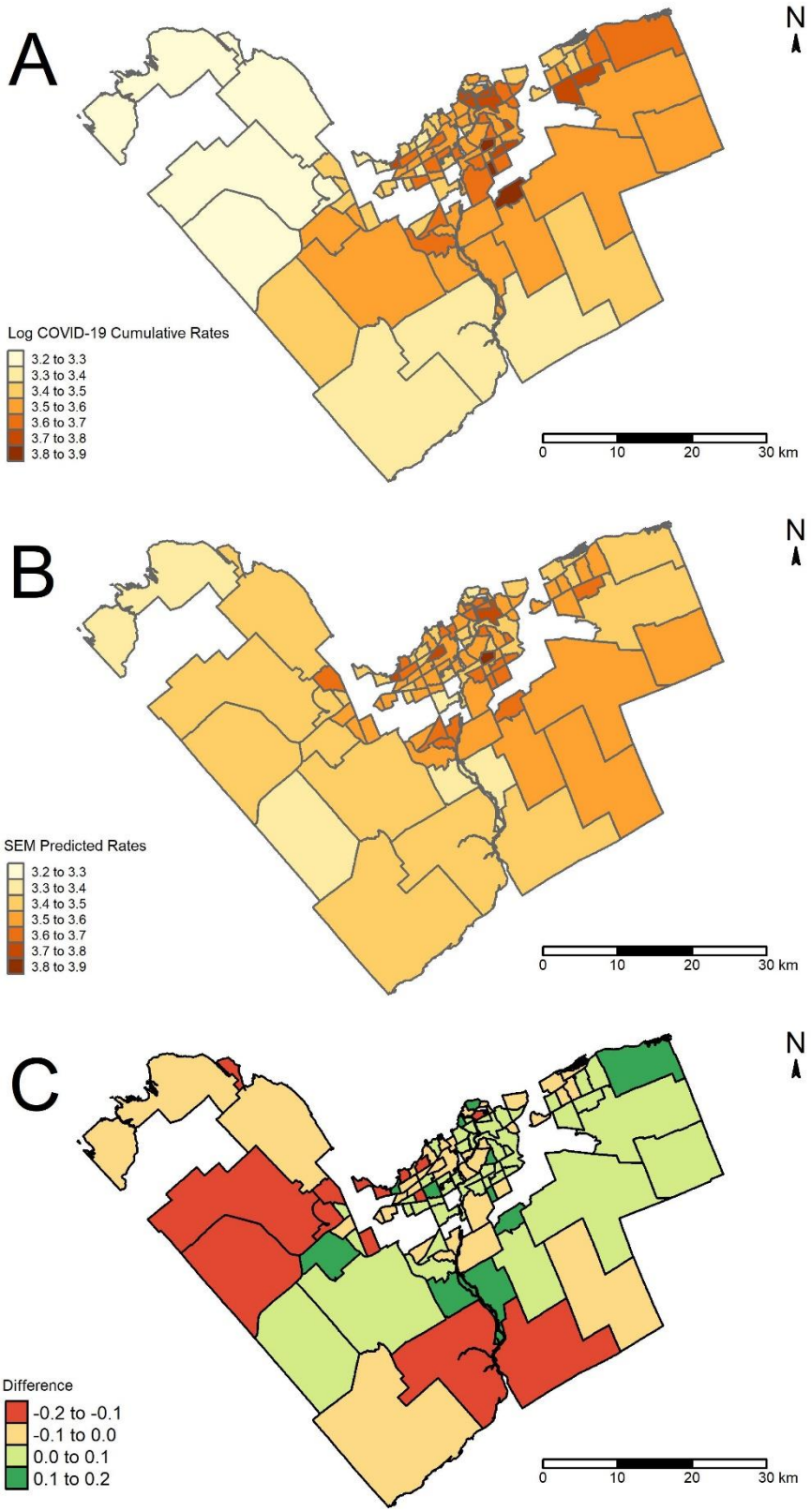


Figure 9. Model Evaluation and Spatial Residual Analysis of COVID-19 Rates in Ottawa, (A) corresponds to the observed log transformed COVID-19 cumulative rates, (B) corresponds to the predicted logged

*transformed COVID-19 rates from the SEM, (C) corresponds to the difference between the observed and predicted SEM COVID-19 Rates (**residual** = observed COVID-19 rate – Predicted COVID-19 rate).*

In general, the northeastern part of the city tends to exhibit underestimation of COVID-19 rates, whereas the northwestern side and southern regions display a tendency to overestimate COVID-19 rates (Figure 9c).

Next, the spatial structure of the SEM residuals was analyzed to determine if the model fully accounted for the bias in the OLS model. A two-sided Moran's I test was performed using the moran.mc() function (Cliff and Ord, 1981) from the spdep package (Bivand 2022) (version 4.3.0). This test indicates that the SEM model has fully accounted for the spatial autocorrelation introduced by the OLS model (Moran's I= -0.04; p = 0.67). Furthermore, the Shapiro-Wilk test resulted in a test statistic that was not significant, providing further evidence that the SEM model is robust, and the residuals are approximately normal with Shapiro-Wilkes test equal to (W=0.98; p = 0.59).

3. The top-down approach to modeling the pattern of COVID-19: Spatial Heterogeneity Analysis

To estimate the optimal bandwidth for the GWR model, The bw.gwr () function from the GWmodel package (version 2.2-9) (Gollini et al. 2015a) was utilized. Next, the kernel specification can be one of Gaussian, Exponential, Bisquare, Tricube, or Boxcar. Each kernel handles weights differently as the distance between data points increases from the center of the kernel bandwidth.

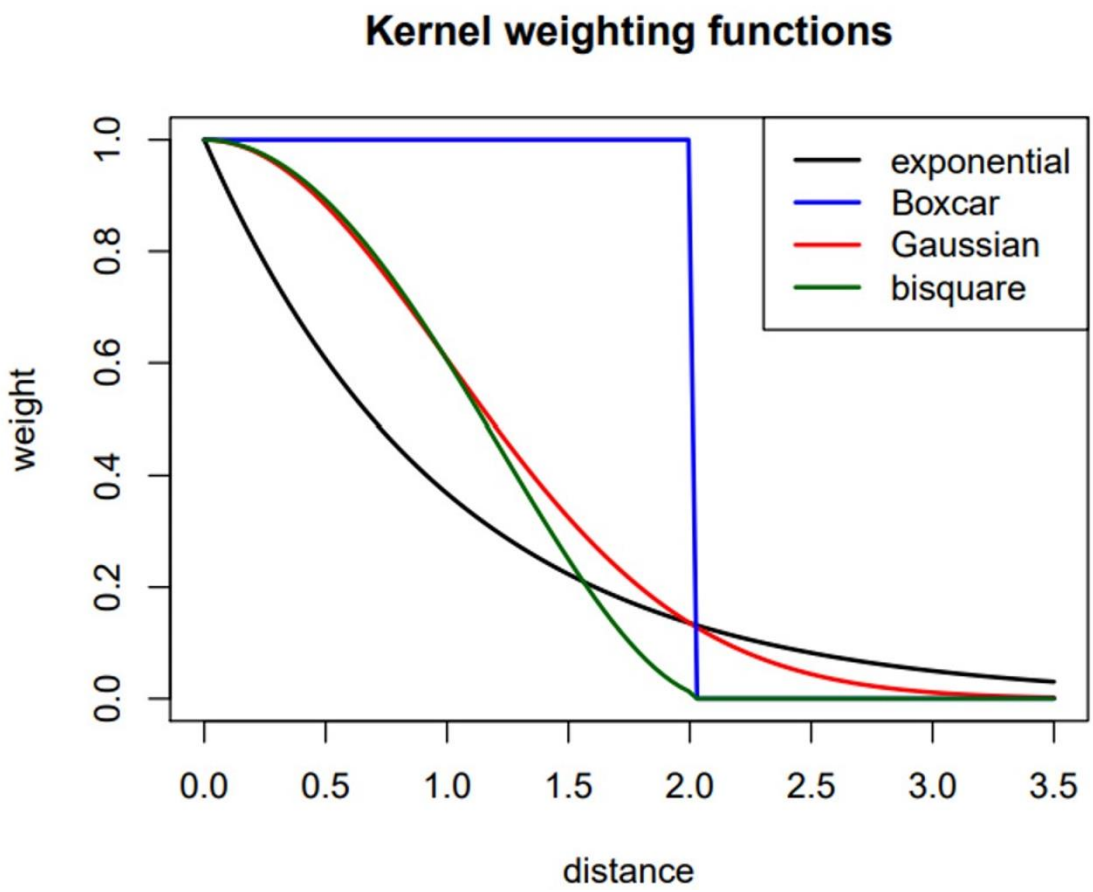


Figure 10. Kernel Weighting Functions for Geographically Weighted Models with Bandwidth of 2.0. Directly from (Rossiter, 2022).

The GWR cross-validation approach produced a fixed bandwidth of 3.4 kilometers for each neighborhood. However, the residuals of this model showed statistically significant positive spatial autocorrelation, and the residuals were not normally distributed, violating the GWR residual assumptions. Subsequently, different combinations of kernel functions with fixed and adaptive bandwidths were examined. However, all residuals displayed positive autocorrelation and/or non-normal distributions, thereby violating the assumptions of GWR. As a result, GWR was unable to consider the varying scales at which COVID-19 rates occurred within Ottawa. Next, multiscale Geographic Weighted Regression (MGWR) was explored (A. Stewart Fotheringham, Yang, and Kang 2017b). In MGWR, each local relationship between COVID-19 and the determinants of the neighborhood of interest was specified using optimal bandwidths for each determinant, allowing for improved modeling of the spatial heterogeneity.

Table 10. Results of Multiscale Geographic Weighted Regression Models Tested with Different Kernel Functions and Bandwidth Options.

ID	Model Description	Adjusted R^2	AIC	Bandwidth	Residuals
1	Kernel Gaussian, Adaptive True	0.67	-238.7	(1)18, (2) 93, (3) 10, (4) 10	Normal distribution, +SA
2	Kernel Exponential Adaptive True	0.66	-246.6	(1)10, (2) 81 (3)10, (4)10	Normal distribution, +SA
3	Kernel Bisquare Adaptive True	0.75	-279.5	(1) 37, (2)100 (3)60, (4)80	Normal distribution No SA
4	Tricube Adaptive True	0.75	-272.4	(1) 41, (2) 100, (3) 67, (4)80	Normal distribution No SA
5	Boxcar Adaptive True	0.72	-247.5	(1) 27, (2) 45, (3) 23, (4)39	Normal distribution, +SA
6	Kernel Gaussian Adaptive False	0.76	-310.7	(1) 4542, (2) 51674.1 (3)2682.5, (4) 1953.3	Not normally distributed. No SA
7	Kernel Exponential Adaptive False	0.74	-418.2	(1)1582.8, (2)51674 (3)2682.4, (4)2552.7	Not normally distributed. No SA
8	Kernel Bisquare Adaptive False	0.71	-252.2	(1)65309, (2)63870 (3)16867, (4)25844	Normal distribution, +SA
9	Tricube Adaptive False	0.70	-249.3	(1)65309, (2)63870 (3)17967, (4)25844	Normal distribution, +SA
10	Boxcar Adaptive False	0.69	-236.4	(1)23935, (2)51674 (3)51674, (4)24404	Normal distribution, +SA

Note: In column Bandwidth, (1) refers to Percentage of people taking public transit to work, (2) Percentage of people with no high school diploma, (3) Percentage of people with bachelor's level degree or above, (4) refers to Percentage of people over 65 years old, +SA refers to a statistically significant positive spatial autocorrelation, when adaptive is true, the bandwidth unit is number of neighborhoods, but when adaptive is false, the bandwidth unit is in meters.

Based on the results of the Shaprio-Wilkes and Moran's I tests, models 3 and 4 (Table 10) were the only two models that met the MGWR assumptions. Both models utilized adaptive bandwidths with a Bisquare and Tricube kernel, respectively (Table 10). Hence, the Bisquare model with an adaptive bandwidth was selected as the ultimate choice for the MGWR analysis, as it has the lowest AIC value (Table 10).

In the selected MGWR model, the number of neighborhoods considered when estimating the regression coefficients for each of the determinants is provided. The bandwidth for each determinant is adaptive and varies depending on the number of neighboring locations. For instance, the bandwidth for the percentage of people who use public transit to work is 37 neighborhoods. This means that the independent determinants of the 37 surrounding neighborhoods were included in the model to estimate COVID-19 rates for each neighbourhood. In contrast, the bandwidth for the percentage of people with no high school diploma is 100 neighborhoods. This implies that the predicted COVID-19 value was most strongly associated by the values of this determinant in the 100 closest neighboring locations, which is almost all neighbourhoods within the entire city of Ottawa, and suggests a global relation for this determinant. Similarly, the bandwidth for the percentage of people with a bachelor's level degree or above is 60 neighborhoods, meaning, the relationship between COVID-19 cumulative rates and the percentage of people with a bachelor's level degree or above varies at a spatial scale of 60 neighborhoods. Whereas it is 80 neighborhoods for the percentage of people over 65 years old. To gain a deeper understanding of the relationship between COVID-19 rates and the determinants above, the estimated regression coefficients across the study area are mapped (Figure 11). By doing so, the spatial variation in the relationship between COVID-19 rates and each determinant is explored.

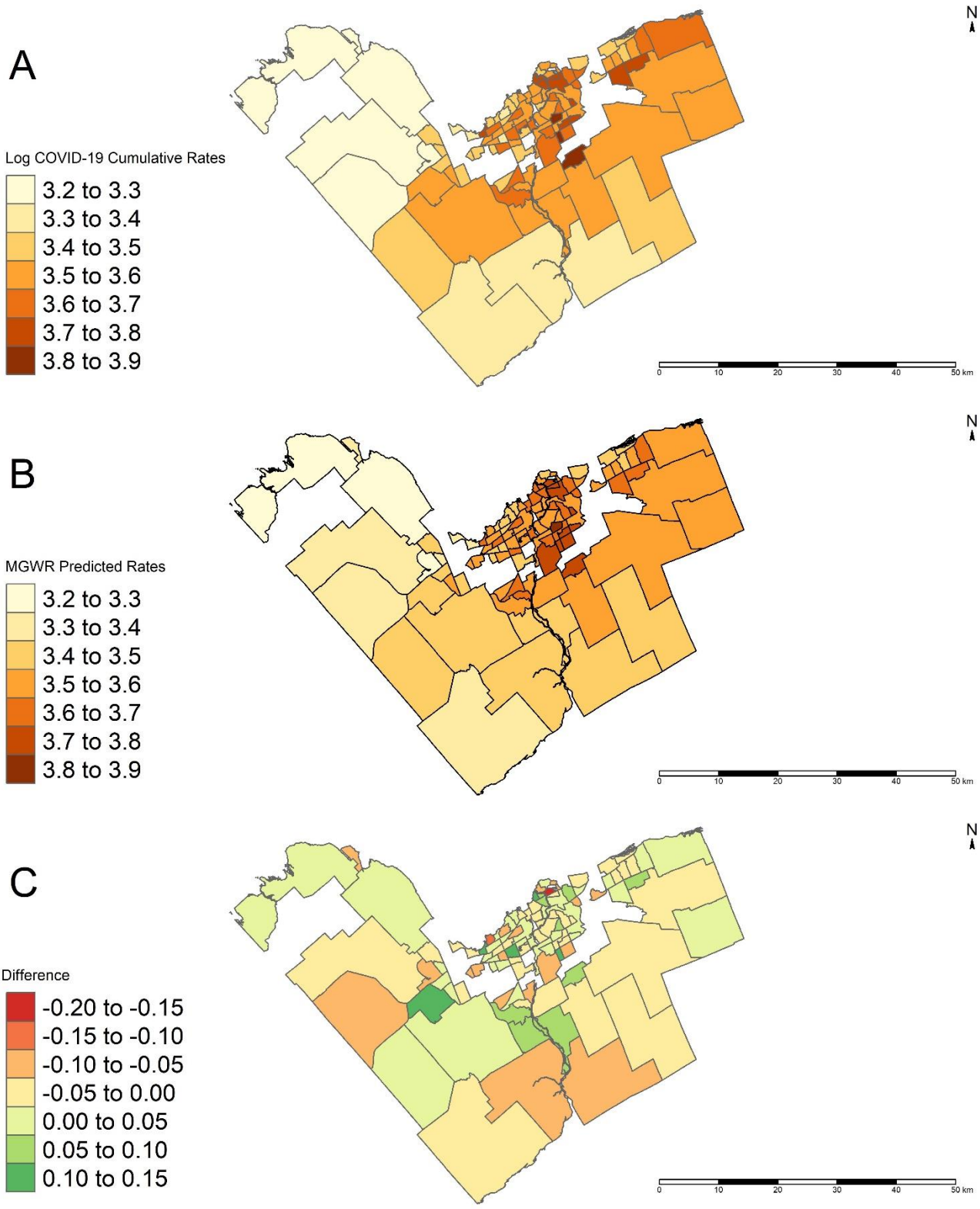


Figure 11. MGWR Evaluation and Spatial Residual Analysis of COVID-19 Rates in Ottawa, (A) corresponds to the observed log transformed COVID-19 cumulative rates, (B) corresponds to the predicted log transformed COVID-19 rates from the MGWR model, (C) corresponds to the differences between the observed and predicted MGWR COVID-19 Rates (Observed - Predicted).

In General, the MGWR model demonstrates a high degree of accuracy ($R^2 = 0.75$) in predicting the actual COVID-19 rates (Figure 11c). The northeastern part of the city consistently reflects accurate predictions of COVID-19 rates, with slight overestimations found in the south. (Figure 11c)

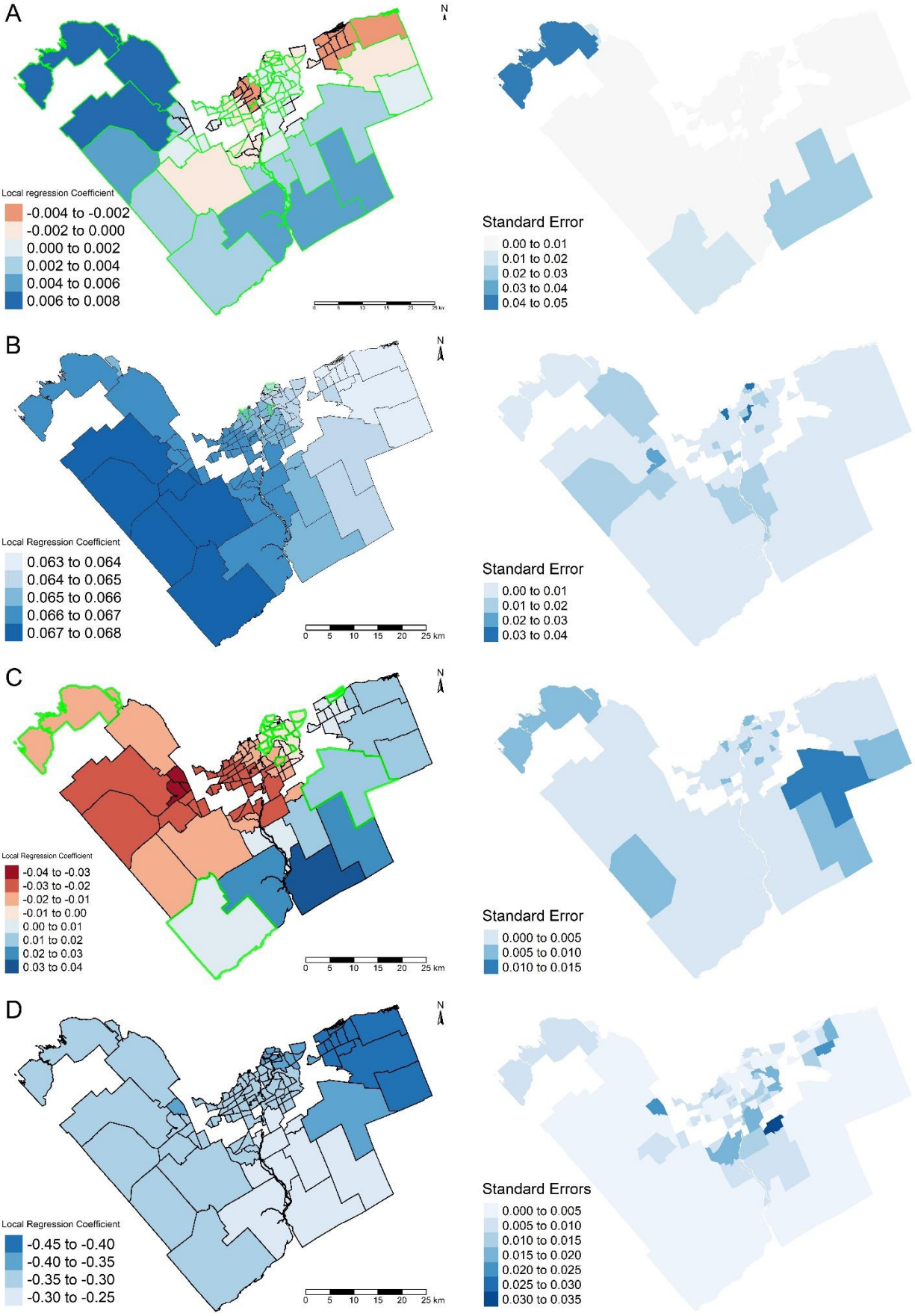

A
N
Local regression Coefficient
-0.004 to -0.002
-0.002 to 0.000
0.000 to 0.002
0.002 to 0.004
0.004 to 0.006
0.006 to 0.008
0 5 10 15 20 25 km
Standard Error
0.00 to 0.01
0.01 to 0.02
0.02 to 0.03
0.03 to 0.04
0.04 to 0.05
B
N
Local Regression Coefficient
0.063 to 0.064
0.064 to 0.065
0.065 to 0.066
0.066 to 0.067
0.067 to 0.068
0 5 10 15 20 25 km
Standard Error
0.00 to 0.01
0.01 to 0.02
0.02 to 0.03
0.03 to 0.04
C
N
Local Regression Coefficient
-0.04 to -0.03
-0.03 to -0.02
-0.02 to -0.01
-0.01 to 0.00
0.00 to 0.01
0.01 to 0.02
0.02 to 0.03
0.03 to 0.04
0 5 10 15 20 25 km
Standard Error
0.000 to 0.005
0.005 to 0.010
0.010 to 0.015
D
N
Local Regression Coefficient
-0.45 to -0.40
-0.40 to -0.35
-0.35 to -0.30
-0.30 to -0.25
0 5 10 15 20 25 km
Standard Errors
0.000 to 0.005
0.005 to 0.010
0.010 to 0.015
0.015 to 0.020
0.020 to 0.025
0.025 to 0.030
0.030 to 0.035

Figure 12. Local MGWR Regression Coefficients For Different Determinants, first column, and Standard Errors second column; (A) Local regression coefficient for percentage of people who take public transit to work, (B) Local regression coefficient for people with no high school diploma, (C) Local regression coefficient for people with a bachelor level degree or above, (D) Local regression coefficient for percentage of people over 65 years old, with statistically significant neighborhoods contoured in black. Note that each row and each map have different maximum and minimum values to highlight the coefficient patterns for each of the beta coefficients. All statistically significant neighborhoods are bordered in a black contour, while non-statistically significant neighborhoods were bordered in green.

The MGWR coefficients for all four determinants exhibited spatial variation across the study region, thus indicating that the relationship between COVID-19 rates and the determinants do vary depending on location. Because of the nature of MGWR as an exploratory rather than confirmatory tool, the the gwr.multiscale() function from the GWmodel package lacks a *p* output, but it does provide *t*-values. Since the large numbers of neighbours (>30) used for each local regression, the *t* distribution would be nearly identical to the standard normal distribution and so these *t*-values were used to compute the *p* values using the pnorm() function (Wichura 1988), from the stats package (R Core Team, 2023, version 4.3.0). By computing twice, the difference between 1 and this probability, a two-tailed *p* was derived for each neighborhood. This resultant *p* can then be utilized to assess the statistical significance of the local regression coefficients for each determinant across Ottawa neighbourhoods. These *p* were then adjusted using the False Discovery Rate (FDR) correction method.

Identifying significant neighborhoods indicates that the observed relationship between COVID-19 and its determinant is unlikely to be a chance occurrence. However, it's important to note that there is a multiple testing issue that has not been accounted for. Thus, the statistically significant MGWR beta coefficients serve as an exploratory tool, offering insights into how the relationships between the determinants and COVID-19 may vary across space (Figure 12). The standard errors map serves to convey the level of uncertainty associated with the coefficient estimates. In instances where the standard error values are high, it's important to be cautious when interpreting apparent significant beta coefficients within different neighborhoods, because the considerable high standard together with multiple testing suggest caution.

For the relationship between the percentage of people using public transit for work and COVID-19 in different neighborhoods, positive coefficients, indicate higher COVID-19 in neighborhoods with greater public transit usage, particularly in the western side of the city (Figure 12a). Conversely, negative coefficients suggest lower COVID-19 rates associated with higher public transit usage, primarily in the northern and eastern sides (Figure 12a). However, it's important to note that these coefficients are generally close to zero and only a few of them are statistically significant, despite very low standard errors. This means that there is limited evidence to suggest a substantial relationship between public transit usage and COVID-19 rates across the city. This lack of substantial variation is further supported by the small standard errors observed in most neighborhoods (Figure 12a). In essence, this variable appears to be relatively stationary, meaning that its impact on COVID-19 rates remains consistent across all neighborhoods, despite this determinant having the smallest scale (37 neighbourhoods) for its optimal bandwidth. The relationship is significant and positive in the SEM model (Table9).

The highest positive coefficients for the percentage of people with no high school diploma are located on the western side of the city, while those with lower positive coefficients decrease towards the eastern part of the city (Figure 12b). The difference in coefficient values is minimal, only 0.05, suggesting a statistically significant but relatively consistent association between a higher percentage of people with no high school diploma and elevated COVID-19 rates across all Ottawa neighborhoods. This consistency is reinforced by the entire study area showing statistical significance except for the three non-significant neighborhoods. This stationary pattern suggests a consistent impact of the percentage of people with no high school diploma on COVID-19 rates across Ottawa. While most neighborhoods display minimal standard errors, the slightly elevated standard errors observed in non-significant neighborhoods may suggest a degree of uncertainty in the coefficient estimates. (Figure 12b).

The percentage of people with a bachelor's degree or above exhibits a subtle spatial pattern in Ottawa. However, the narrow coefficient range of just -0.04 to +0.04 (0.08 units) suggests a relatively modest association that doesn't strongly indicate elevated risk. While most neighborhoods are statistically significant, exceptions in the northern part of the city highlight variations in the impact of this determinant on COVID-19 rates. It's also important to note that many non-significant neighborhoods came out with slightly higher standard errors. This suggests that in these areas, there is more uncertainty in the coefficient estimates. In other words, the influence of the percentage of people with a bachelor's degree on COVID-19 rates in these neighborhoods may be less certain (Figure 12c).

The negative relationship between the percentage of people over 65 years old and COVID-19 rates across Ottawa is noteworthy. It suggests that neighborhoods with a higher percentage of seniors tend to have lower COVID-19 rates (Figure 12d). The range of coefficients, spanning from -0.45 to -0.25, underscores the strength of this inverse association. This negative relationship is particularly pronounced in the northeastern part of Ottawa, where it has a more substantial impact on COVID-19 rates. In contrast, the inverse negative relationship is weaker in the south and western side of the city, indicating that age has a lesser impact on COVID-19 rates in those areas. While some neighborhoods did not show statistical significance, it's important to note that the majority did, emphasizing the influential role of the percentage of people over 65 years old in explaining the variation in COVID-19 rates across Ottawa. This impact is most prominent in the northeastern side of the city. However, it's worth mentioning that neighborhoods with slightly higher standard errors for the percentage of seniors' determinant tend to cluster in the central part of the city and just south of the Ottawa Greenbelt, with Findlay Creek displaying the highest standard error. These areas exhibit a degree of uncertainty in the coefficient estimates, suggesting that while age plays a significant role in explaining COVID-19 rate variations, there may be other factors at play in these specific neighborhoods (Figure 12d).

Next, the residuals of the MGWR model were examined to verify the assumption of normality. It is essential to assess whether they follow a normal distribution or not.

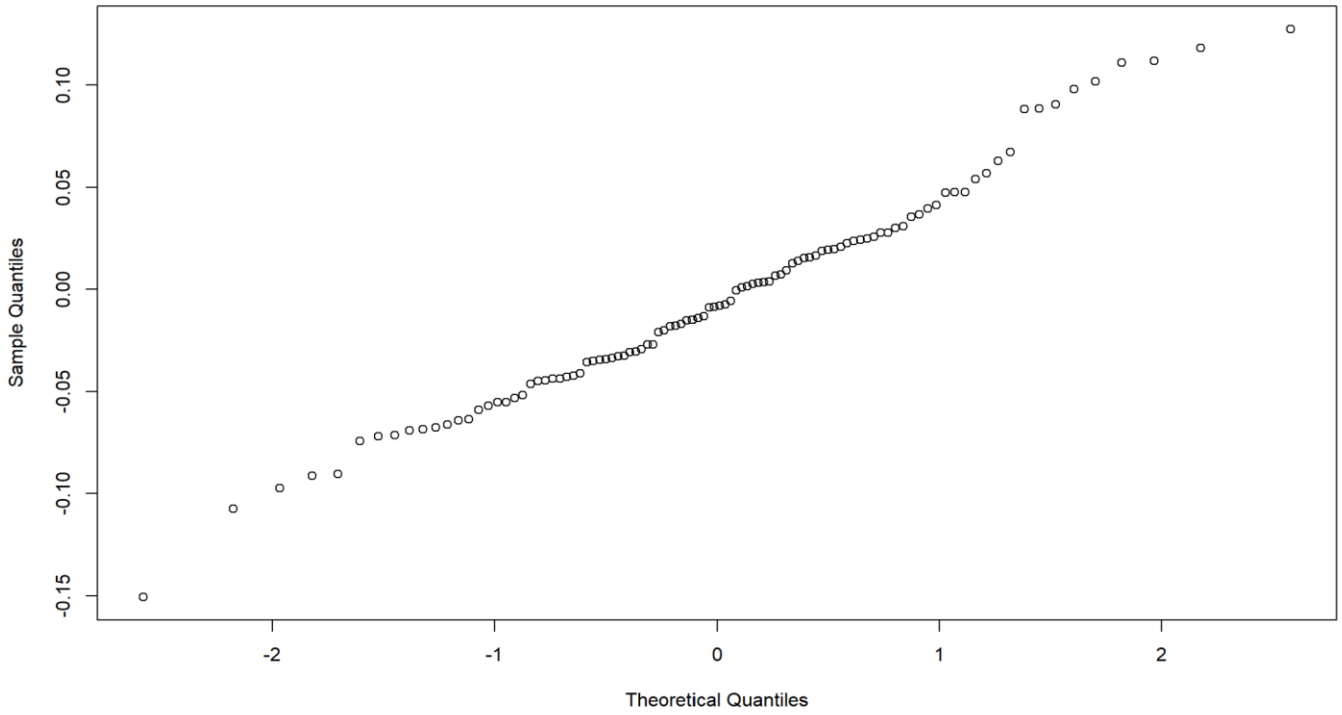


Figure 13. QQ plot of the MGWR residuals.

With only a small deviation, the residuals closely follow a normal distribution, evident from the nearly perfect straight diagonal line (Figure 13). Furthermore, the Shapiro-Wilk test resulted in a test statistic that was not significant, providing further evidence for the approximate normal distribution of the MGWR's residuals ($W = 0.98$; $p > 0.25$). Also, a two-sided Moran's I test was performed using the moran.mc() function and indicated that there is no significant spatial autocorrelation in the residuals ($I = -0.04$; $p = 0.62$). The pattern of residuals across Ottawa does not exhibit any clustering or spatial structure beyond what would be expected by random chance. Both the SEM exhibited and MGWR residuals lack spatial structure.

4. The bottom-up approach to modeling the pattern of COVID-19: Data Mining using Random Forest

To prepare the dataset for the Random Forest algorithm, the dataset was refined by selecting 238 socioeconomic determinants, out of an initial list of 356, in a total of 102 neighborhoods (Appendix 3). This variable selection step ensured uniformity in the representation of all determinants, accommodating various units. This step prevents determinants with absolute counts from impacting the results since the focus was to use standardized variables, for example, number of supermarkets per person not number of supermarkets only. This helps the algorithm to make equitable comparisons across all features.

Subsequently, the dataset was divided into training and testing subsets, using the R (R Core Team, 2023, version 4.2.2) built-in sample () function (Becker, Chambers & Wilks, 1988). This function aided in randomly selecting rows for the training set, with the remaining rows allocated to the testing set. Specifically, 80% of the data formed the bootstrap, or the training set, while the remaining 20% constituted the testing set. This approach ensures that the training set is diverse and representative of the overall dataset, enabling the model to learn from various patterns present in the data. The testing set however, provides a different measure for evaluating the model's predictive performance/generalization ability on data that the model has not evaluated.

To identify the optimal number of trees for the random forest model, values ranging from 1 to 350 trees were tested using the randomForest() function (Breiman 2001b) from the randomForest package (version 4.7-1.1) (Liaw and Wiener 2002). The loop trained a total of 350 random forest models. However, it is important to note that these models were used solely for training purposes, and the testing data was reserved for a final and unbiased evaluation of the selected model. Subsequently, these models were utilized to predict COVID-19 cumulative rates on the testing data. By comparing the predicted values with the actual values, the mean squared error (MSE) was computed. The MSE quantifies the average squared difference between the predicted and actual values, thereby indicating the model's performance. The MSE values were stored, and the minimum MSE was identified to determine the random forest model with the optimal number of trees. In this case, it was found that the model achieved the lowest MSE when using 18 trees, suggesting that this configuration provides the best performance in predicting COVID-19 cumulative rates in Ottawa.

Building upon the same methodology used to determine the optimal number of trees, a similar iterative process was conducted to identify the optimal number of determinants (maximum number of features) at each split. By iterating through values ranging from 1 to 238, representing the total number of determinants in the training/testing set, random forest models were built. Utilizing the training data, the models were trained and subsequently used to predict values on the testing data. Again, the MSE was then computed and the optimal value, which leads to the most accurate predictions, was determined by identifying the minimum MSE value. In this case, the analysis revealed that 83 was the optimal value, providing the best performance in estimating COVID-19 cumulative rates. The rest of the arguments were set to default.

The model accounted for a substantial portion of the variability in COVID-19 cumulative rates, as evidenced by the 47.34% of variance explained.

In a random forest analysis, it is critical to evaluate the variability of residuals against normality assumptions. This evaluation is vital for establishing the reliability and validity of the model. However, for this study, which focuses on 102 neighborhoods in Ottawa, with 81 in the training set and the remaining 21 in the testing set, the available number of residual points is limited to 21. Consequently, the small sample size poses challenges for conducting spatial autocorrelation tests, leading to limitations in forming definitive conclusions.

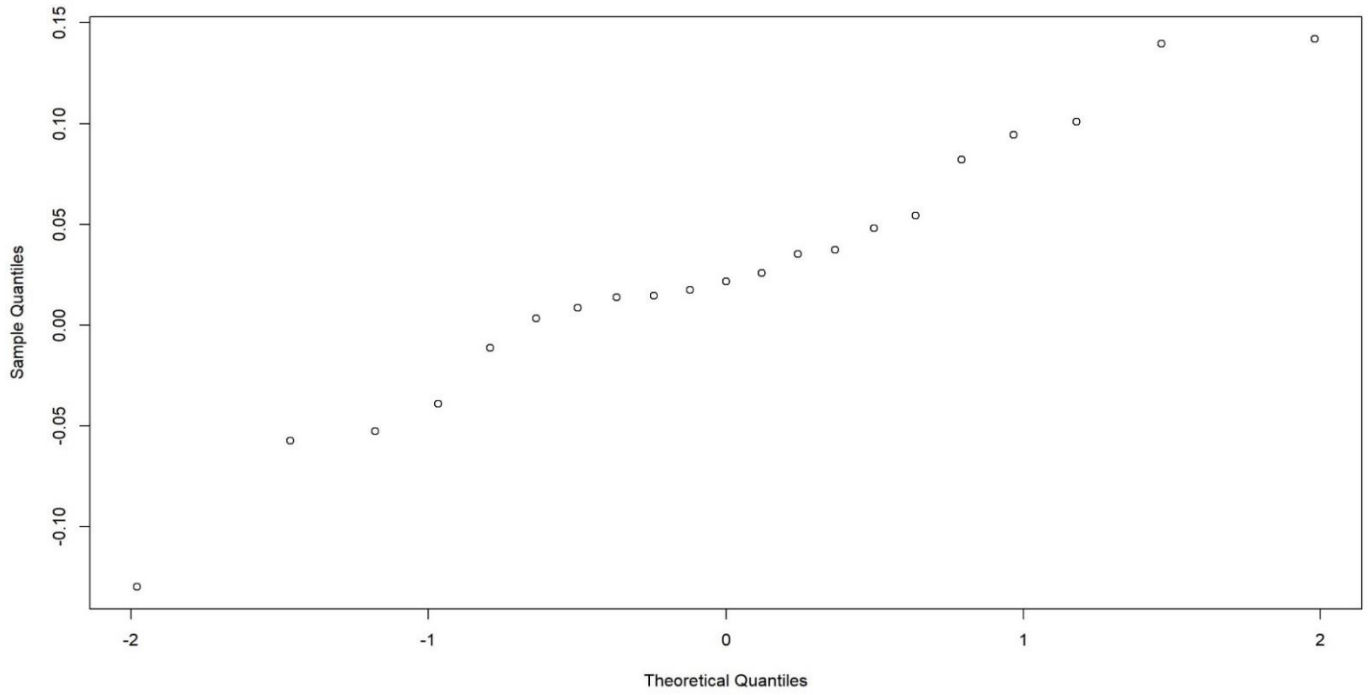


Figure 14. QQ residual plot

An almost perfect straight diagonal line indicates that the distribution of the residuals closely follows a normal distribution. The random forest model's residuals exhibit a distribution that is approximately normal (W = 0.96; p= 0.68) (Figure 14). The p represents the probability of obtaining the observed test statistic (or a more extreme value) if the null hypothesis is true. In this case, the null hypothesis is that the random forest residuals follow a normal distribution. With a p=0.68, there is insufficient evidence to reject the null hypothesis. Based on the Shapiro-Wilk test result, along with the earlier observation of an almost perfect straight diagonal line in the QQ plot, there is reasonable evidence to support the assumption of normality for the random forest residuals.

Next, to evaluate the spatial patterns generated by the Random Forest model, a prediction map of the log transformed COVID-19 cumulative rates was created.

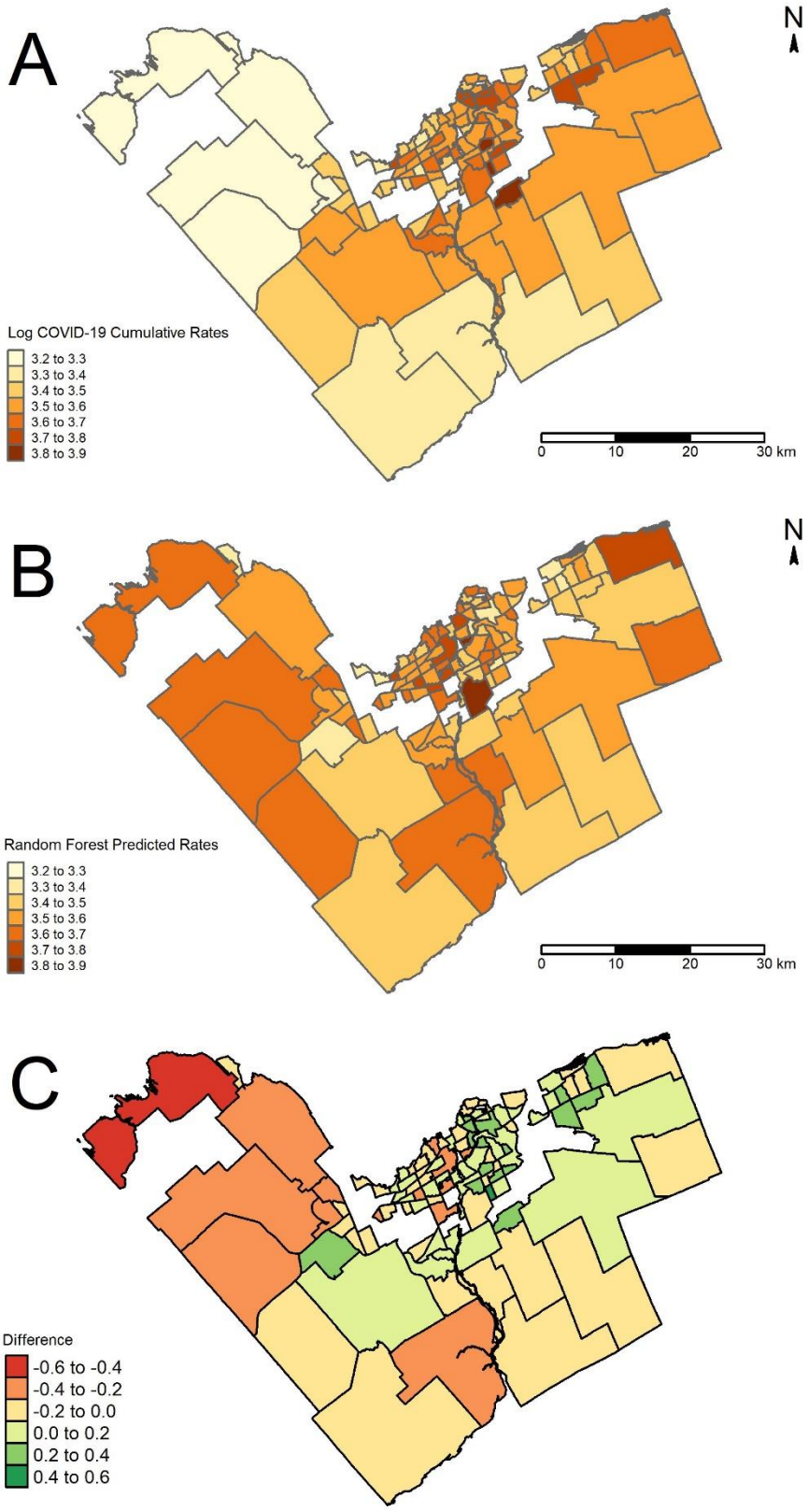


Figure 15. Comparison of logged transformed COVID-19 rates, (A) corresponds to the log transformed observed rates, (B) corresponds to the Random Forest prediction map, (C) corresponds to the difference between the observed and the Random Forest predicted COVID-19 log transformed rates.

Visually, the Random Forest model did not achieve the same level of prediction accuracy as its preceding models, the SEM and MGWR. This is apparent due to the cluster of overestimations in the northern parts of the city (Figure 15c).

One of the primary advantages of employing random forest is its ability to identify the most important features that impact COVID-19 rates in Ottawa. In the previous multivariate regression analyses, the percentage of people over 65 years old, the percentage of people with a bachelor's level degree or above, the percentage of people with no high school diploma, and the percentage of people who take public transit to work were identified as influential determinants of COVID-19 rates based on the literature. The objective is to compare the determinants identified through regression analysis, based on the literature, with the determinants identified by the random forest approach.

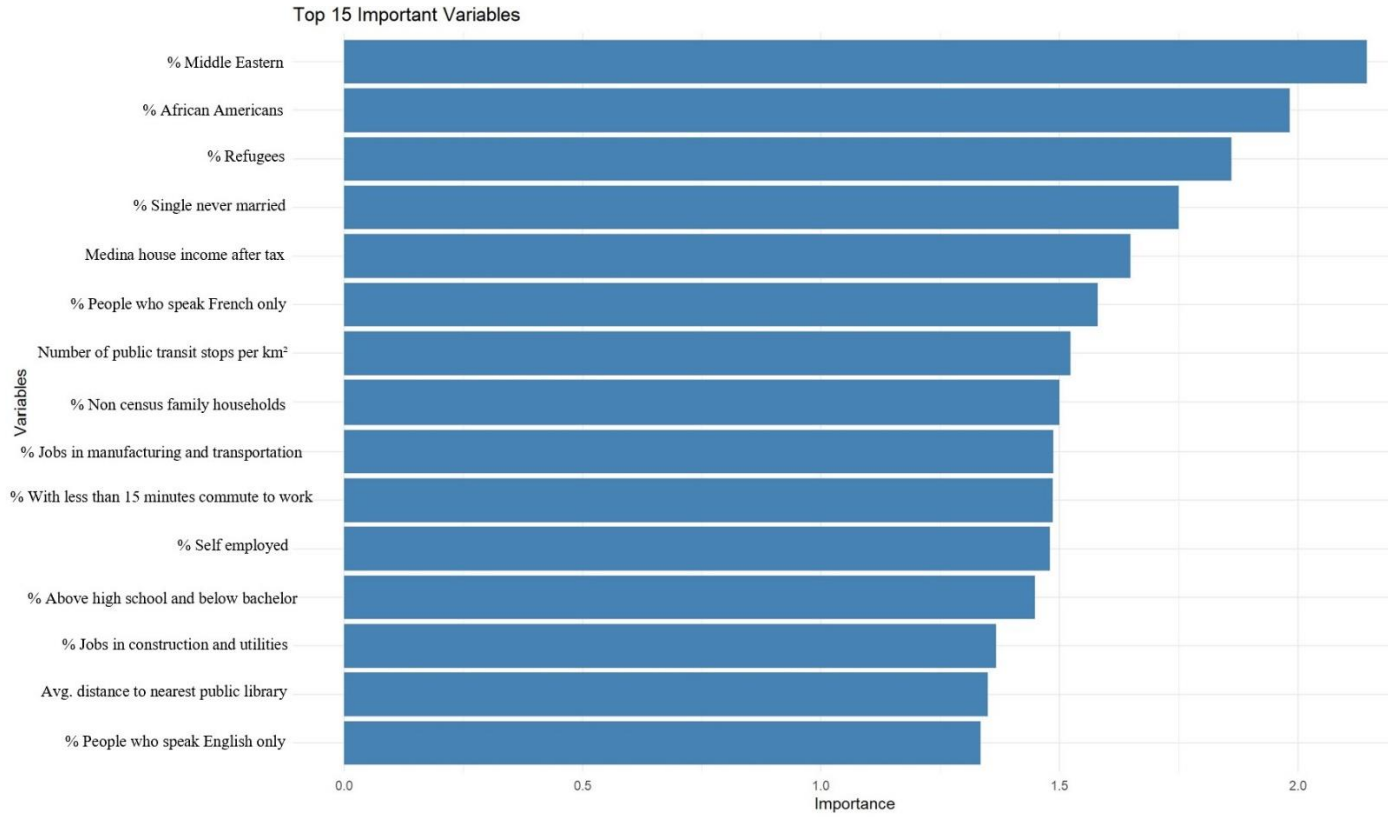


Figure 16. Top 15 Important determinants: Insights from Random Forest Analysis.

The determinants resulting from the random forest analysis can be categorized into four distinct subgroups, offering insights into the factors associated with COVID-19 rates in Ottawa (Table 11).

Table 11. Categorized Random Forest Determinants Influencing COVID-19 Rates in Ottawa.

Subgroup	Determinants
Ethnicity and Demographic	Percentage of Middle Eastern, Percentage of African Americans, Percentage of refugees.
Socioeconomic	Median income after tax, Percentage of people who speak French only, Percentage of people who speak English only, Percentage of non-census family households, Percentage of single never married, Average distance to nearest public library.
Transportation	Number of public transit stops per km², Percentage with less than 15 minutes commute to work
Occupation	Percentage of jobs in manufacturing and transportation, Percentage of jobs in construction and utilities, Percentage of self-employed, Percentage of above high school and below bachelor's.

Model Comparison

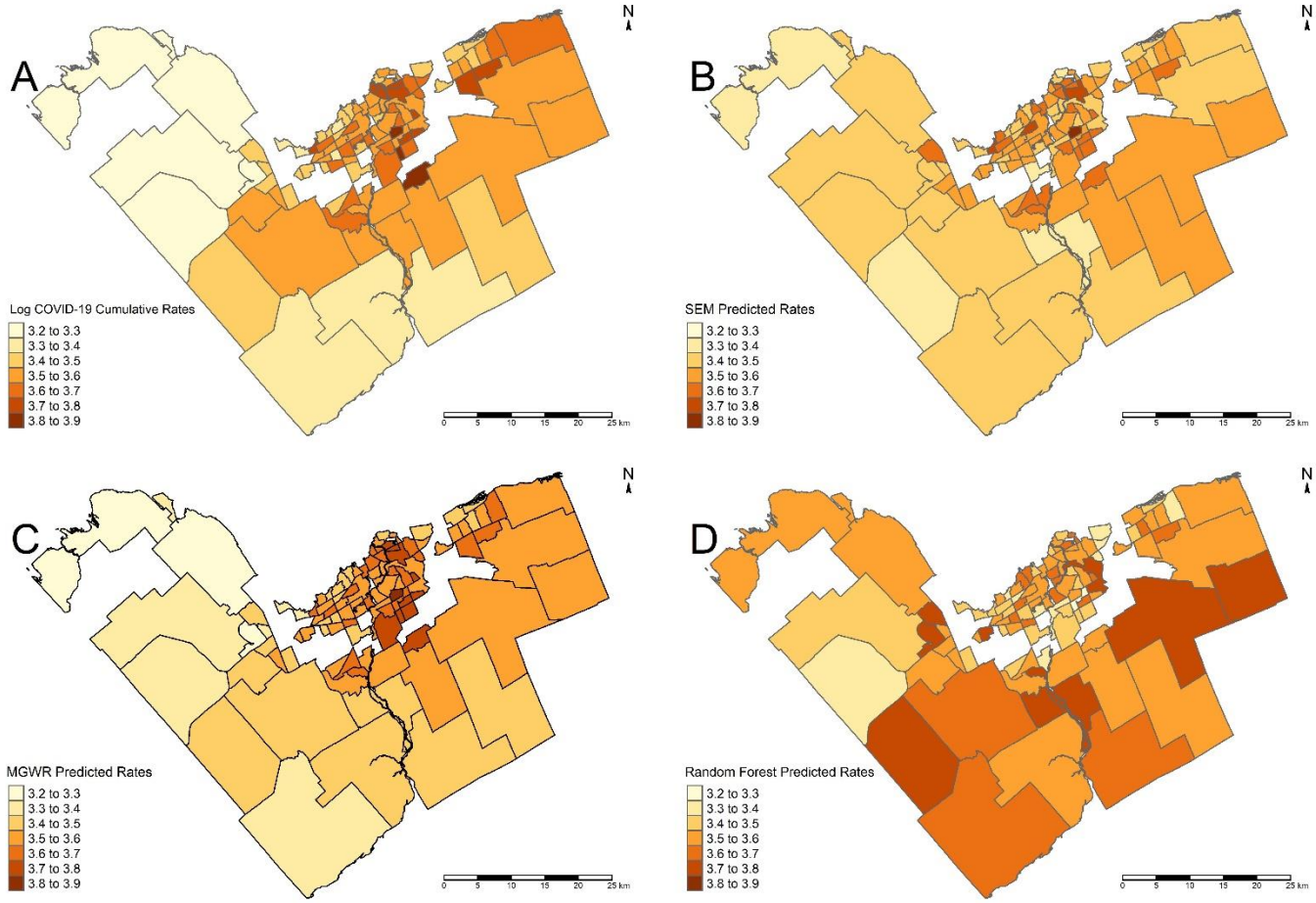


Figure 17. (A) corresponds to the observed COVID-19 cumulative rates with transformation applied, (B) corresponds to the SEM predicted logged transformed COVID-19 rates, (C) corresponds to the MGWR predicted logged transformed COVID-19 rates, (D) corresponds to the Random Forest predicted logged transformed COVID-19 rates. Visually, the MGWR prediction map excels in accuracy when forecasting COVID-19 rates, surpassing the SEM prediction map, which tends to underestimate in the northeast and overestimate in the northwest (Figure 17).

An alternative method for comparing models, beyond relying on visual assessment, involved the application of Lee's test. Lee's test is a spatial bivariate correlation statistic, designed to evaluate the spatial correlation existing between different models. This

assessment takes into account the spatial relationships among the models' predictions. To conduct this analysis, the lee.test() function (Lee 2004a) from the spdep package (R. Bivand 2022) was employed.

Table 12. Table of Lee's L Test Results.

	Observed	SEM	MGWR	Random Forest
Observed		0.33*	0.5***	0.12
SEM	0.44***		0.35*	0.04
MGWR	0.5***	0.48***		0.13
Random Forest	0.12	0.06	0.13	

Note: * indicates ($p \leq 0.05$) and *** ($p \leq 0.01$)

The MGWR model emerges as the strongest performing model, by its moderate statistically significant spatial correlation between the predicted and the observed COVID-19 ($L = 0.5$). This suggests that the MGWR model's predictions are closely aligned with the observed values in terms of their spatial arrangement. The SEM also shows a weak statistically significant bivariate spatial correlation ($L = 0.33$) (Table 12).

Chapter 4: Discussion

1. Spatial Dependence Analysis in the pattern of COVID-19 in Ottawa: Global and Local Moran's I

The global spatial dependence of COVID-19 rates in Ottawa was statistically significant and moderately positive spatial autocorrelation (SA) and confirms that the pattern of the disease is not random (I= 0.39; p = 0.0001). The presence of spatial clustering implies that neighborhoods with high COVID-19 rates tend to be close to other neighborhoods with high rates, while neighborhoods with low rates tend to be close to other neighborhoods with low rates, as is the case with most spatial data (Tobler 1970). Similar statistically significant positive spatial autocorrelation patterns of COVID-19 have been identified in relevant studies from various regions across the world (Cordes and Castro 2020; Vilinová and Petrikovičová 2023).

Numerous studies have employed the Local Moran's I method to identify statistically significant patterns and clusters within their datasets, allowing for a more nuanced exploration of differences and disparities in COVID-19 infection rates across different geographic areas. Local Moran's I helps researchers pinpoint areas of similarity or dissimilarity in their regional analyses, revealing pockets of nonstationary and individual locations that differ from the global statistic. By doing so, it enables a deeper understanding of the variations in infection rates across different neighborhoods or regions, guiding the allocation of resources and interventions where they are most needed (Cordes and Castro 2020; Ghosh and Cartone 2020).

The local Moran's I analyses also identified two cold-spot neighborhoods, namely, Borden Farm - Fisher Glen and Hunt Club Woods - Quintarra, situated in the northern part of the city (Figure 5) COVID-19 rates lower compared to their immediate neighbors. To gain a deeper insight into the determinants contributing to these lower local COVID-19 rates, an examination was conducted on some of the determinants based on the top 15 most important determinants, identified by the random forest model (Figure 16).

The analysis was expanded to encompass the immediate neighbors of Hunt Club Woods, which include Ottawa Airport, South Keys, Hunt Club East, Riverside Park, Billings Bridge, and Ledbury. Similarly, for Borden Farm, the immediate neighbors considered were Carleton Heights, Fisher Heights, Cityview, Tanglewood, and Parkwood Hills. This extension allowed for a detailed comparison of values between these spatial outliers and their neighboring areas. It's worth noting that Greenbelt and Hunt Club South (industrial neighbourhood), which are also immediate neighbors of these outliers, were excluded from the analysis due to the absence of COVID-19 infection rates data. To provide a benchmark for comparison, the entire dataset average was used for comparison.

Table 13. Determinants Characteristics in Cold Spot Neighborhoods (blue) and Surrounding Areas(red) with Ottawa (green).

		% Middle Eastern	% Black	% Refugees	% Single Never Married	# Public Transit Stops Km²	% of Jobs in Manufacturing, Transportation	% of Jobs in Construction, Utilities	% Seniors
	Ottawa	4.5	6.6	4.3	30.4	2.1	8.4	4.2	14.4
	Borden Farm	4.3	2.3	2.9	26.5	17	0	3.8	26.6
	Hunt Club Woods	4	4.7	3.9	23.1	9.7	3.6	2.1	27.6
Hunt Club Woods cluster	Ottawa Airport	8.4	5.4	2.2	25.4	3.2	31.5	0.3	7.7
	South Keys	12.4	7.2	4.6	32.7	16.7	6.2	9.6	19.9
	Hunt Club East	12	13.1	8.5	34.3	18.4	0.4	7.6	15.8
	Riverside Park	6.7	3.9	11.6	30.6	14.9	37.4	0.1	20.0
	Billings Bridge	5.1	5.5	3.9	33.4	20.0	2.4	2.0	18.6
	Ledbury	15.1	27.4	22.8	41.7	17.9	7.1	19.2	8.9
Borden Farm cluster	Carleton Heights	5.7	8.4	5.3	42.1	15	0.5	4.3	15.7
	Fisher Heights	3.6	6.8	3.9	28.7	14.5	1.4	2.1	20.5
	City view	5.2	2.5	2.7	36.5	12.9	1.4	1.9	18.8
	Tanglewood	5.1	8.8	5.2	33.4	13.2	4.3	2.1	17.6
	Parkwood Hills	13.2	16.3	11.8	46.1	41.2	4.1	3.6	9.5

When using the Local Moran's I, the focus is on local relationships rather than global ones. In this context cold spot rates are lower than those of their immediate neighbors and are, generally lower than the mean of the entire dataset.

The cold spot neighborhoods mostly exhibit lower rates across all determinants compared to Ottawa, which represents the mean, with one exception being the percentage of seniors. However, the findings from the SEM model (Table 9) unveiled an inverse relationship between COVID-19 rates and the percentage of seniors. This regression coefficient explains the higher rates for this determinant within the cold spot neighborhoods, aligning with the SEM model's findings.

In Hunt Club Woods, several trends emerge when comparing its demographic and socioeconomic indicators with those of its neighboring areas (Ottawa Airport, South Keys, Hunt Club East, Riverside Park, Billings Bridge, Ledbury). Notably, the percentage of Middle Eastern residents consistently remains lower in Hunt Club Woods compared to all surrounding neighborhoods. Similarly, the percentage of Black residents is consistently lower, except for Riverside Park, which exhibits a higher percentage. Furthermore, the percentage of single, never married people consistently registers lower values in Hunt Club Woods when contrasted with its immediate neighbors. These patterns persist across various determinants, with some exceptions. For example, when examining the percentage of jobs in the Manufacturing and Transportation sector, Billings Bridge and Hunt Club East stand out with lower percentages compared to Hunt Club Woods (Table 13).

For Borden Farm, the percentage of Middle Eastern residents is generally higher in all neighboring areas (Carleton Heights, Fisher Heights, City view, Tanglewood, Parkwood Hills), except for Fisher Heights. In contrast, for Black residents, this percentage is consistently higher in all surrounding neighborhoods. While these patterns hold for most determinants, there are a few exceptions, where the rates at cold spots are slightly higher than the immediate neighbors, particularly in determinants related to occupation (Table 13).

It's important to note that the percentage of seniors in Borden Farm and Hunt Club Woods follows a different trajectory, displaying higher values in comparison to their nearby neighborhoods. This anticipated disparity can be attributed to the negative relationship between the percentage of seniors and COVID-19 rates (Table 13).

Identifying outliers, including cold spots, holds significant practical implications for policymakers and researchers. The key lies in understanding the context of these outliers in relation to a reference point, such as a model regression line. When an outlier falls above the regression line, it signals areas that exhibit higher-than-expected values. In such cases, this knowledge can prompt immediate attention and resource allocation, as these areas may be potential hotspots requiring targeted interventions. Conversely, when an outlier falls below the regression line, indicating a cold spot, it implies that a specific area has achieved lower-than-expected rates. These areas offer valuable insights into effective practices or community dynamics that have successfully reduced infection rates. Researchers and policymakers can examine these low-spot outliers in the rates in order to uncover reasons that could lead to best practices, community-based initiatives, or unique strategies that have resulted in lower rates.

These observations are particularly relevant for Kanata Lakes, a neighborhood with a lower COVID-19 rates. Kanata Lakes and surrounding region was put on the top priority list for COVID-19 vaccines by the Ontario government which surprised regional experts (Figure 5) (Perez, 2021). What perplexed both residents and observers alike was the unexpected designation of Kanata Lakes as a COVID-19 hot spot on the provincial list to receive the vaccine. Although the health minister defended this choice as justified, the government has not yet provided the methodology supporting this decision, leaving Kanata Lakes classified as a vaccine priority without evidence or clarity (Perez, 2021).

In contrast, Ottawa Public Health, recognized for its transparent methodology, identified various neighborhoods with higher rates, excluding Kanata Lakes from that list (Perez, 2021). On the flip side, Bay Ward, characterized by a significant population in subsidized housing and a high number of essential workers, exhibited a combination of determinants traditionally signaling the need for vaccine priority (Statistics Canada, 2023). However, Bay Ward found itself excluded from vaccine prioritization at the provincial level in favor of Kanata Lakes (Perez, 2021).

Speculation has emerged that individuals with Ontario government authority residing in Kanata Lakes may have used their authority to include their neighbourhood(s) as vaccination priorities, potentially gaining early access to vaccinations. This suspicion is fueled by the government's failure to publicly announce the methodology used in classifying Kanata Lakes' status as a hot spot, and in need of a vaccine priority (Members of Parliament, n.d.). To assess the legitimacy of this decision, future research could

replicate the analytical approach used in this study using COVID-19 monthly rates, up to the point when Kanata Lakes was classified as a hot spot, providing a comprehensive evaluation of the decision's justification.

2. The top-down approach to modeling the pattern of COVID-19: Multivariate Regression Analysis and uncovering Determinants Contributing to COVID-19 Rates in Ottawa

The analysis of residual plots from the OLS backwards stepwise regression indicated Kanata Lakes, Britannia Village, Briarbrook, and Constance Bay, as potential outliers (Figure 6). These outliers represent data points that deviate from the expected values predicted by the regression model. These deviations could be attributed to unique local determinants, specific demographic characteristics, that are associated with lower COVID-19 rates.

Table 14. Neighborhood Transformed Determinants for the Outliers and Areas with Positive Spatial Autocorrelation in Low-Low COVID-19 Rates.

	Neighborhood	Population Density	Job Density	Newcomers	Income	Public Transit	Seniors	COVID-19 Rates
Outliers	Constance Bay	19.4	1.1	0.0	36212.0	3.0	1.1	3.3
	Kanata Lakes	37.1	2.7	2.1	46643.0	16.5	1.2	3.2
	Briarbrook	40.9	3.4	2.0	41803.0	17.5	0.9	3.4
	Britannia Village	54.9	2.9	2.2	29761.0	34.0	1.4	3.4
+SA in the L-L rates	Beaverbrook	48.0	2.4	1.8	38159.0	17.0	1.3	3.4
	Katimavik	45.4	3.1	1.4	38548.0	20.3	1.1	3.5

+SA in the L-L rates presents positive spatial autocorrelation in the low-low rates.

Beaverbrook and Katimavik have been identified as neighborhoods showing positive spatial autocorrelation in the low COVID-19 rates (Figure 5). These neighborhoods were included alongside the outliers to analyze some of their determinants (Table 8). Comparing the determinants of the outliers to the low COVID-19 rates neighborhoods may provide an explanation for the shared low COVID-19 rates (Table 14).

Constance Bay has the most pronounced deviation from the entire dataset's average, notably including a lower population density, reduced dependence on public transit for commuting, lower percentage of newcomers, and lower job density. Also, Constance Bay exhibits comparable income levels, similar percentages of senior residents, and low rates of newcomers when compared to Beaverbrook and Katimavik (Table 14). As a result, Constance Bay also exhibits lower cumulative COVID-19 rates (Table 8). While this study found a positive correlation between population density and COVID-19 rates (Figure 4), other studies found that higher population densities tend to have lower COVID-19 mortality rates, possibly owing to the presence of more advanced healthcare systems in densely populated areas. (Hamidi, Sabouri, and Ewing 2020)

Lower population density, higher income, and lower job density (Table 8) in Kanata Lakes all exhibit bivariate positive correlations with COVID-19 rates, except for increasing

income which is associated with a decrease in COVID-19 rates (Figure 4). Also, Kanata Lakes exhibits rates that are quite similar to most determinants in Beaverbrook, apart from a lower population density and higher income (Table 14). Moreover, despite the controversy regarding the inclusion of K2V, a postal code within Kanata Lakes, as a COVID-19 hotspot (McKay, 2021), it was one of the first neighborhoods in Ottawa to become eligible for COVID-19 vaccines, as it was deemed a high-priority area (Williston, 2021). Altogether, these determinants are associated with lower COVID-19 rates in this neighborhood in comparison to the entire city of Ottawa.

Briarbrook exhibits comparable rates to Katimavik in terms of population density, job density, and the percentage of senior residents. Its rate of newcomers is similar to Beaverbrook, and it shows a higher income level than both of these neighborhoods (Table 14). This similarity of determinants with other neighborhoods that are positively autocorrelated in the low COVID-19 rates, along with a high income, can explain its low COVID-19 rates compared to the city's average.

In Britannia Village, two factors stood out distinctly: job density and the percentage of residents over 65 years old (Table 8). The bivariate correlation between COVID-19 rates and the percentage of people over 65 years old is negative, implying that neighborhoods with a higher proportion of elderly residents tend to have lower COVID-19 rates. Job density exhibits a positive bivariate correlation, as previously mentioned (Figure 4). Hence, with higher-than-average percentage of senior residents and lower job density, it suggests that these determinants may contribute to the low COVID-19 rates in this neighborhood. Also, Britannia Village shares greater similarity with Beaverbrook concerning population density, job density, and the percentage of senior residents. In contrast, when it comes to income level, Katimavik exhibits closer resemblance (Table 14).

Three out of the four potential outliers, namely Kanata Lakes, Constance Bay, and Briarbrook, exhibit positive spatial autocorrelation and are characterized by statistically significant low COVID-19 rates (Figure 5). This spatial autocorrelation indicates that these neighborhoods are geographically clustered in a way that aligns with their lower COVID-19 rates. Such clustering suggests the presence of shared local factors or conditions that might be contributing to the lower rates.

These neighborhoods have unique characteristics, such as lower population density, higher income, lower job density and potentially other determinants that collectively create a protective effect against COVID-19. These findings are consistent with existing research on the effects of income inequality (Demenech et al. 2020) and population density (Wong and Li 2020) on COVID-19 outcomes.

The SEM model provided coefficients for each determinant (Table 9). Specifically, a one-unit increase in the square root of the percentage of people with no high school diploma equates to 10.6% increase in COVID-19 rates. A one-unit increase in the square root of the percentage of people with a bachelor's degree or above is associated with a 3.2% increase in COVID-19 rates. A one unit increase in the percentage of people using public transit to work is associated with a 0.1% increase in COVID-19 rates. Conversely, a one percent increase in the percentage of people aged 65 and older is linked to a 34.6% decrease in COVID-19 rates.

Hence, the inverse relationship associated with the percentage of seniors and COVID-19 rates could be attributed to older individuals having a heightened awareness of their vulnerability to the virus, perhaps due to their awareness of contagion given experience in avoiding annual influenza. It is also possible that seniors, due to their age and living arrangements, may not interact as extensively with the general population that, through working conditions perhaps have higher exposure to the virus, such as those working in manufacturing and service industries. This reduced interaction could act as a protective factor for seniors, lowering their risk of exposure to COVID-19. As a result, they are more inclined to prioritize their health and seek prompt medical advice, thereby contributing to lower infection rates.

The results of this study contrast with those of Mansour et al. (2021), whose findings indicate a positive association between increasing elderly population (aged 65 and above) and COVID-19 incidence rate. Mansour et al. (2021) observed that an increase in the size of the elderly population was linked to an elevated rate of disease incidence, with increasing elderly population being the most significant determinant. It's worth noting that Mansour et al. (2021) doesn't specify whether cases linked to retirement homes and outbreaks in long-term care facilities were excluded, whereas in this study they were, and this information gap underscores the importance of clarity and transparency in research methodology for accurate comparative analysis and interpretation. This difference is also explained by how Omanis, like many Muslim families in the world, do not tend to have large numbers of long-term care facilities for elderly family members and instead, the elderly stay within the family unit all living in one household. If a member of the household gets COVID-19 then the more vulnerable elderly members will most certainly succumb to the illness. In the present research, had long-term facilities been included in the analyses, there may have been an opposite relation between aged 65 and older and COVID-19 rates. It is well-known that rates were much higher in such facilities (Government of Canada 2021; Vilches et al. 2021) and this would have added considerably to the rates in the containing neighbourhoods.

On the other hand, the positive relationship between cumulative COVID-19 rates and the percentage of people with no high school diploma, people with a bachelor’s level degree or above and percentage of people who take public transit to work could be explained by a few different factors. Individuals who rely on public transit may feel a heightened risk of virus exposure due to their proximity to fellow travelers during their journeys. The increased risk prompted a drop in ridership for Ottawa-Carleton Transportation (OCTranspo) (Raymond, 2021). Many individuals, realizing the prudence of working from home as a precautionary measure against COVID-19, chose not to use public transportation when they no longer had to commute to work due to lockdown measures. Moreover, the lack of significant neighbourhoods in the MGWR analysis suggests a disconnect between public transit use and COVID-19, that is to some degree, effected by the fact that the census determinants were all collected during normal non-pandemic lockdowns. Therefore, the relationship found in the analysis may be spurious and due to some unknown confounding variable, that effects both public transit usage during normal times and COVID-19.

Percentage of people with no high school diploma have a positive relationship with COVID-19 rates (Figure 5). This correlation can be influenced by a range of factors, such

as disparities in educational achievement and consequent socio-economic conditions. These factors can impact attitudes towards vaccination and perceptions of COVID-19. Research indicates that vaccine hesitancy is significantly associated with respondents who have lower education levels and rely predominantly on social media for COVID-19 vaccine information in Jordan, Kuwait, and other Arab nations (Sallam et al. 2021). As indicated by the random forest analysis (Figure 16), immigrants from Middle Eastern origins emerged as the most significant determinant. This prominence suggests that beliefs in COVID-19 vaccine-related conspiracy theories, such as the idea that vaccines intend to inject microchips for control or are linked to infertility, could be an explanatory factor in the importance of the Middle Eastern determinant concerning COVID-19 rates. This significance may arise if a considerable portion of the Middle Eastern diaspora in Ottawa holds culturally driven hesitancy towards accepting the vaccine (Sallam et al. 2021). It's important to note that while social media reliance is linked to vaccine hesitancy in Arab countries, the emphasis on Middle Eastern origins in this analysis is a distinct finding.

Considering the MGWR model, the east-west division in relation to COVID-19 local regression coefficients based on the percentage of people with a bachelor's degree or higher is not very pronounced, as the differences hover close to zero (Figure 12c). French speakers exhibit a positive bivariate correlation with COVID-19 rates (Figure 19). It's worth noting that the eastern part of Ottawa has a higher concentration of Francophone populations compared to the western part (Figure 18). However, other studies found that multiple factors can be associated to this division other than educational attainment such as gender, vaccination acceptance and income. (Allington et al. 2023)

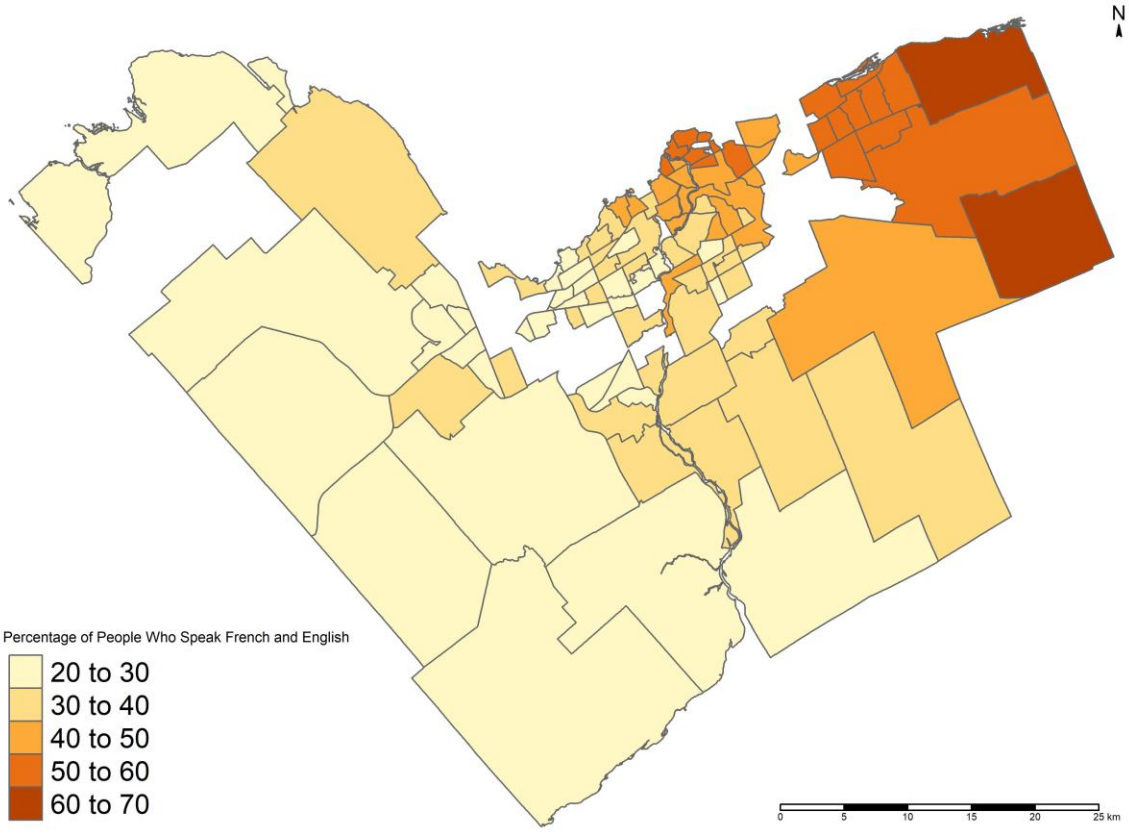


Figure 18. Distribution of the Percentage of People who Speak French and English in Ottawa.

Supporting this inverse relationship between COVID-19 rates and the percentage of people with a higher educational attainment level on the western side of the city, Sallam et al. (2021) found that potential factors associated with higher rates of vaccine acceptance included male sex, higher educational levels, and a history of chronic disease in Kuwait, Jordan, and other Arab nations. This study was chosen for discussion in the context of the analysis because the random forest model identified immigrants from Middle Eastern origins as the most important variable influencing COVID-19 rates in this study area, despite the Middle East region not being culturally close to Canada. Furthermore, with respect to the primary source of COVID-19 vaccine information, individuals who relied on medical doctors, scientists, and scientific journals exhibited lower levels of belief in vaccine-related conspiracies (Sallam et al. 2021). In another study in Italy, Zarbo et al. (2022) found that vaccine hesitancy is not limited to educational levels but also encompasses individuals who are relatively young, face economic hardships, express heightened psychological distress due to the pandemic, and are particularly concerned about potential vaccine side effects. (Zarbo et al. 2022)

Higher educational attainment was found to be negatively correlated with COVID-19 rates in other studies (Siljander et al. 2022; Han et al. 2021). However, this present study revealed a significant positive but very weak relationship between a bachelor's level degree or above and COVID-19 rates on the eastern side of the city and the opposite on the western side when looking at the MGWR coefficients (Figure 12c). Several potential determinants could contribute to the positive relationship observed in eastern Ottawa.

These include factors such as increased testing and reporting, occupational disparities exposing educated individuals to in-person interactions, and socioeconomic influences that may link education level with international travel opportunities. This connection with international travel could potentially introduce new cases of the virus. In conclusion, the influential role of determinants such as public transit ridership, educational attainment, and age on COVID-19 rates was found to be evident.

3. The top-down approach to modeling the pattern of COVID-19: Multiscale Geographically Weighted Regression

The decision to choose the Multiscale Geographic Weighted Regression (MGWR) model over the Geographically Weighted Regression (GWR) model was based on several factors. Firstly, the residuals of the GWR model exhibited statistically significant positive spatial autocorrelation, indicating that model failed to capture the spatial dependency in the data. Secondly, the non-normal distribution of the GWR residuals violated an underlying assumption of the GWR model. These limitations suggested that the GWR model does not properly capture the relations between the COVID-19 rates in Ottawa and the independent variables. Therefore, the MGWR model was adopted. The MGWR model offers enhanced capabilities by allowing for localized relationships to be specified using adaptive bandwidths, that is varying neighbourhood sizes for each determinant, thus providing a more accurate representation of spatial variability of the relation between COVID-19 rates and determinants (A. Stewart Fotheringham, Yang, and Kang 2017b).

The MGWR model's higher adjusted R^2 value (0.75; Table 9) compared to the SEM (0.63; Table 8) suggests that MGWR is more effective at explaining the variation in COVID-19 rates. This is likely due to MGWR's ability to capture spatial heterogeneity, adapt to local effects, and handle spatial autocorrelation, which are valuable when analyzing any spatial patterns (Fotheringham et al., 2017). These findings align with prior research that also found MGWR to be more effective in explaining COVID-19 rates in comparison to other global spatial regression models (Mansour et al. 2021; Shi et al. 2023).

Local regression coefficients for public transit usage for work and COVID-19 rates exhibit positive coefficients in the south and west of the city, indicating increasing COVID-19 rates in those neighbourhoods as public transit utilization increases, while negative coefficients in some of the central and northeastern neighborhoods imply the opposite association (Figure 12a). This east-west dichotomy contrasts with expectations of greater transit usage equating to higher COVID-19 rates because of people's proximity within buses, which could facilitate disease transmission.

It's important to note that the public transit data used in this analysis pertains to the period before the COVID-19 pandemic. The assumption is that, keeping all other factors constant, the patterns of ridership would remain consistent across Ottawa due to the implementation of mitigation measures. Consequently, one might expect higher COVID-19 rates in areas with greater ridership. However, it's crucial to acknowledge that reductions in ridership during the COVID-19 pandemic may not follow a uniform proportionality across different neighborhoods. A study has shown that this variation is linked to the diverse job profiles of essential workers in Ottawa (Government of Canada, Statistics Canada, 2022).

Ottawa experienced lockdown measures in response to the COVID-19 pandemic (Crawford, 2021). The decreased dependence on public transit among Ottawa residents had a major impact on public transit usage patterns (Raymond, 2021). Buses operating below full capacity, a restricted number of individuals using them for essential work commutes, and a shift in travel preferences likely played a role in reducing the potential for widespread virus transmission on public transportation. Moreover, it appears that Ottawa had a well-organized hygiene strategy in place, involving frequent surface sanitization, human resources increase, and enforced preventive measures to mitigate the impact of public transit on COVID-19 transmission (Free Transit Ottawa - Covid and the Public Sector, 2020).

There is a consistent near-zero or negative local regression coefficient relationship between all the determinants and COVID-19 rates on the western side of the city (Figure 12). Additionally, the same area of the city exhibits local positive spatial autocorrelation in the low COVID-19 rates (Figure 5).

The local regression coefficients for the percentage of people without a high school diploma are all positive and statistically significant. These coefficients suggest that areas with a higher concentration of residents lacking a high school education, particularly in the western part of the city, tend to have a higher prevalence of COVID-19 (Figure 12b). The same neighborhoods characterized with high COVID-19 rates due to lower educational attainment are also positively locally spatially autocorrelated with lower COVID-19 rates in nearby neighborhoods (Figure 5). In simpler terms, it means that areas with more residents lacking a high school diploma are linked to higher COVID-19 rates and are part of the western region’s large cluster of low-low COVID-19 rates when compared to the overall city average.

One might initially expect a negative correlation between the percentage of people with a bachelor's level degree or above and COVID-19 rates (Figure 4). This is certainly the case for most neighbourhoods in western Ottawa with the negative coefficients predominantly found in the western and central areas of the city (Figure 12c). Conversely, positive coefficients are concentrated in the eastern side of the city (Figure 12c). In essence, neighborhoods with higher educational attainment tend to exhibit higher COVID-19 rates. This means that in western neighborhoods, where there are significantly lower COVID-19 rates (Figure 5), the relationship is reversed. Here, an increase in higher education is associated with lower disease rates, contrasting with the trend where increasing ridership and lower high school completion equate to higher COVID-19 rates. This finding suggests that higher education may act as a protective factor against COVID-19 at a population level when there are other factors working against a neighbourhood (lower educational attainment, greater use of public transit, and race) (Hawkins, Charles, and Mehaffey 2020). At the same time, in absence of those other factors, positive local regression coefficients for the bachelor’s and above in the east indicate that those neighborhoods with more highly educated residents might experience higher COVID-19 rates. One possible explanation could be that areas with a higher concentration of highly educated individuals might also exhibit greater mobility, international connectivity, or other socio-economic factors that inadvertently contribute to increased virus transmission.

When examining the negative regression coefficients between the percentage of individuals aged 65 and above and COVID-19 rates across various neighborhoods across Ottawa (Figure 12d), it becomes imperative to account for several potential confounding variables, especially given the numerous studies that demonstrate a positive correlation between the percentage of elderly individuals and COVID-19 rates (Péterfi et al. 2022; Li et al. 2020). This is because it is counterintuitive that the elderly population, who are generally more susceptible to illness, would have a negative correlation with COVID-19 rates. This seemingly contradictory result warrants a careful examination of other factors that may be influencing this relationship.

To further reduce susceptibility among senior citizens, public health officials implemented more targeted interventions. For instance, specific hours in grocery stores were designated for senior citizens to reduce interactions with younger individuals who may be carrying the virus (Raymond, 2020). It's important to note that data from long-term care facilities was not considered in this study. Nevertheless, despite attempts to safeguard senior citizens in the community, long-term care facilities in Ontario witnessed a notable increase in mortality rates among their residents (Akhtar-Danesh et al. 2022).

Ottawa's extensive public health initiatives aimed at protecting both its general population and, more specifically, its senior citizens during the COVID-19 pandemic have been tremendous. However, it's crucial to recognize that the negative correlation between the percentage of seniors and COVID-19 rates, as depicted in Figure 4, necessitates further exploration. The existing comprehension of this relationship relies on multiple determinants, including healthcare infrastructure, vaccination coverage, and the implementation of public health interventions, which have not been comprehensively addressed within the scope of this study.

4. The bottom-up approach to modeling the pattern of COVID-19: Random Forest and the most important Determinants.

The random forest analysis conducted in this study identified several subgroups of influential factors impacting COVID-19 rates in Ottawa. These subgroups include ethnicity and demographic factors, socioeconomic factors, transportation patterns, and occupational characteristics. Each subgroup offers insights into the potential explanations for the observed variations in COVID-19 rates.

To evaluate the strength and direction of the relationships among the top 15 variables identified through the random forest analysis, a correlation matrix was employed (Figure 19).

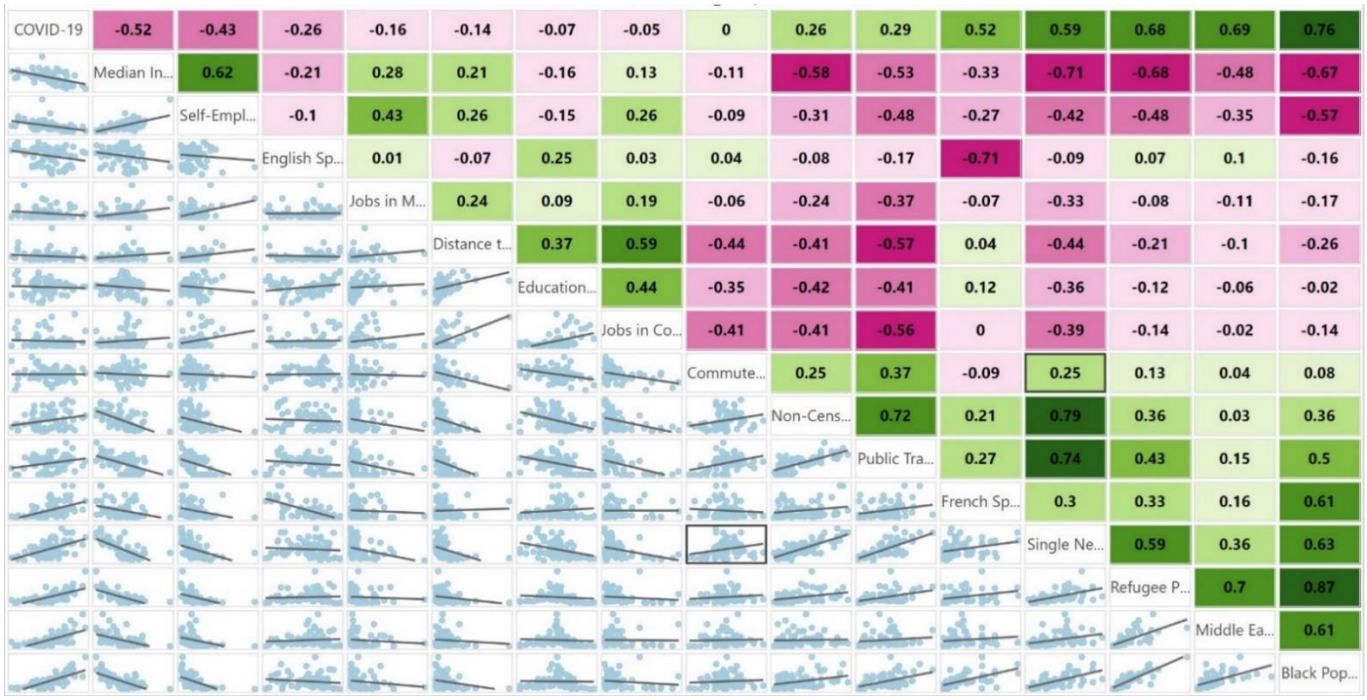


Figure 19. Correlation Matrix of the top 15 variables resulting from the Random Forest model and COVID-19 cumulative rates with the Log transformation applied.

Table 15. Variable Abbreviations Reference for Figure 19.

Variable	Description
Black pop	Percentage of black people
Middle Ea	Percentage of immigrants from Middle Eastern origins
Refugee P	Percentage of refugees
Single Ne	Percentage of individuals who are single and have never been married
French Sp	Percentage of exclusive French speakers
Public Tra	Number of public transit stops per km²
Non-Cens	Percentage of non-census family households
Commute	Percentage of individuals with a commute time of less than 15 minutes to work
Jobs in Co	Percentage of jobs in construction and utilities

Education	Percentage of individuals with education above high school and below bachelor's
Distance t	Distance to the nearest public library
Jobs in M	Percentage of jobs in manufacturing and transportation
English Sp	Percentage of exclusive English speakers
Self-Empl	Percentage of self-employed individuals
Median In	Median income after tax
COVID-19	Cumulative COVID-19 rates with log transformation

Higher COVID-19 rates have a strong bivariate correlation with a higher Percentage of Black residents, a higher Percentage of Middle Eastern residents, and a higher Percentage of refugees in Ottawa, while simultaneously displaying a high negative correlation with median income after tax (Figure 19). However, the Percentage of Middle Eastern immigrants and Percentage of refugees have a very weak positive bivariate correlation with English speakers only and strong negative relationship with median income, supporting that middle eastern immigrants and refugees in Ottawa, tend to face greater language barriers, potentially hindering their social integration and access to job opportunities (Figure 19).

Canada committed to provide a safe home for refugees through various initiatives, such as the Refugee and Humanitarian Resettlement Program (Government of Canada, 2022). Also, Canada's popularity as a destination for high-quality education is also evident, as Canada has experienced significant growth in the number of international students choosing to study in the country (Government of Canada, 2022). For new immigrants, of lower socioeconomic status or refugees with nothing, the government program is particularly significant given the rising cost of living in Canada (Rogers, 2023). The relevance lies in the fact that, due to their commitment to establishing a foothold in the country, newcomers may find themselves working in environments (waste management, taxi driving, custodial services etc..) that expose them to higher risks of COVID-19 infections. This increased exposure is a factor to consider in understanding the dynamics of the pandemic within this demographic. In other words, neighbourhoods with significant immigrant populations are mostly of a lower socioeconomic status and may have higher rates of COVID-19.

The percentage of Middle Eastern population emerged as the most important variable from the random forest model (Figure 16). This finding is significant because it highlights the link between newcomers' mistrust in the healthcare system in their country of origin and vaccine hesitancy. In many countries in the Middle East, there is a prevailing atmosphere of mistrust in the government, limited transparency, and challenges to democratic

governance (Warf, 2015). It is plausible that these mistrust values carry over to the Middle Eastern experience in western countries. (Valero-Martínez et al. 2023)

Households classified as non-census families—those deviating from the traditional census definition of a family unit, including individuals living independently or with unrelated individuals like roommates or unrelated cohabitants—demonstrate a positive bivariate correlation with COVID-19 rates. This correlation aligns with the percentage of single, never-married individuals and suggests common underlying factors, such as an elevated risk of virus transmission within the household and larger social networks with frequent interactions beyond the immediate family (Figure 19). Moreover, non-census family household have a strong relationship with the number of public transit stops per km², suggesting more diverse living arrangements and an increased demand for public transportation services (Figure 19). Notably, a study in Ontario, Canada has emphasized the importance of social distancing as one of the most effective measures in mitigating transmission by lowering the daily COVID-19 infection contact rate (Wu et al. 2020).

Given the negative correlation between income and COVID-19 (Figure 19), individuals with lower incomes often face resource constraints that can affect their access to healthcare and the ability to implement essential preventive measures. This economic disparity results in reduced access to testing, healthcare services, and personal protective equipment, increasing the vulnerability of lower-income individuals to COVID-19. (Elgar, Stefaniak, and Wohl 2020)

One potential underlying factor related to the percentage of people who speak French only in Ottawa, a predominantly English-speaking city, could be the presence of recent newcomers from French speaking origins. The strong positive relationship between the percentage of black population and the percentage of French speakers only further supports this hypothesis (Figure 19). While recognizing that some Canadians in Ottawa are bilingual and speak both English and French, the observation here suggests that individuals who exclusively speak French may be more likely to be newcomers to the city. This assumption takes into account the linguistic diversity within Ottawa, where residents may choose to primarily speak French even if they are not recent arrivals. However, access to French speaking doctors and nurses is less in Ottawa compared to English speaking primary care. Thus, there may also be a hesitancy to seek care and become informed about the benefits of preventative medicine like vaccination.

Since most sub-Saharan African countries were colonized by France, their residents are considered native French speakers (Ager 2005). These newcomers may have recently arrived in the city and have yet to learn English. As a result, language barriers and their status as newcomers, potentially associated with low income, could contribute to a positive relationship with COVID-19 rates.

While the relationship between COVID-19 and the percentage of only French speakers is moderately positive, the same relationship with the percentage of people who are English-only is weak but negative. English speakers represent a more established segment of the population in Ottawa with potentially better language proficiency and comprehension of public health guidelines, as opposed to non-English speakers, specifically many immigrants, who may not fully understand the protocols and guidelines and may not have as much total access to health resources (Clark et al. 2020). The French language and

COVID-19 rates are by no means clear, and further analyses could examine native French-Canadian speaker density vs. non-native francophones and see if there is a difference in either variable's relation to COVID-19 rates. That was however outside of the scope of this research and would require custom tabulations from Statistics Canada. However, that type of analysis would clarify any conjectures made here.

There is a moderately positive relationship between population density and COVID-19 rates (Figure 4). This suggests that areas with higher population density face an increased risk of COVID-19 transmission and this has been found in other studies (Rocklöv and Sjödin 2020). Similarly, a higher percentage of individuals with a commute time of less than 15 minutes suggests a more localized workforce and potentially shorter distances between residential areas and workplaces. The potential mechanism here is that often with reduced travel times and the high concentration of workplaces in the central core of Ottawa, workers usually live in highly dense housing regions within a series of high-rise apartments, that lead to high population density and that are close to work. In such cases, these are often lower-income housing, and close to amenities within walking distance where higher degrees of exposure can take place. In contrast, the more affluent upper middle class tend to live in suburbs in single family homes or small-rise apartments that lead to lower population density and higher travel distances to work. Perhaps more importantly, as an airborne virus, high density regions most often contain large apartment complexes where air is recycled, and neighbours share common hallways and that can increase the risk of exposure and, all other things being equal, inflates the number of cases within a neighbourhood via direct contagion. COVID-19 and population density are closely linked, as research globally has consistently shown (Bhadra, Mukherjee, and Sarkar 2021). In densely populated areas, the virus spreads more easily due to close person-to-person contact in workplaces, public transportation, and social settings. Additionally, healthcare systems in such areas can become overwhelmed, making it challenging to provide adequate care. Enforcing preventive measures like social distancing can also be more difficult, and socioeconomic factors further complicate the situation. As a result, densely populated regions often experience higher infection rates and face unique challenges in managing and mitigating the impact of COVID-19 (Rocklöv and Sjödin 2020).

Lastly, the occupation group comprises determinants such as percentage of jobs in manufacturing and transportation, percentage of jobs in construction and utilities, percentage of self employed, and percentage of above high school and below bachelors. Occupations in manufacturing, transportation, construction, and utilities involve physical presence at a workplace, close contact with others, leading to a higher exposure risk. Surprisingly, the relationship between these determinants and COVID-19 is nearly zero (Figure 19). This implies that while these occupations may carry a higher exposure risk due to their close person to person proximity, the data doesn't show a strong linear correlation between them and COVID-19 rates, but they may play important roles in the non-linear RF model. However, as workers had the opportunity to take sick days due to COVID-19 because the virus was regarded as an unsafe work hazard, they could still receive payment during their absence, thanks to government-provided benefits (Canada Revenue Agency, 2023). Along with the lockdown measures that considerably decreased road traffic and the concentration of workers in one location, this could elucidate the nearly

non-existent association between COVID-19 and the occupation determinants. (Pamidimukkala and Kermanshachi 2021)

On the other hand, self-employment provides more flexibility and may not require in-person attendance, showing a moderate negative relationship with COVID-19 and reducing exposure risks (Figure 19). Research has indeed demonstrated that remote work can be an effective measure in flattening the curve during a pandemic, by reducing person-to-person interactions. However, it's crucial to recognize that remote work is not feasible for all sectors of the workforce (Mustajab et al., 2020). Lastly, the level of education, particularly those above high school and below bachelor's level, indicate better health literacy, influencing adherence to public health measures. Conversely, lower education levels create barriers to comprehending and adhering to healthcare guidance, potentially increasing susceptibility to the virus.(Sallam et al. 2021)

Chapter 5: Limitations and Conclusion

1. Limitations

In this research, as with any ecological level analysis, the Modifiable Areal Unit Problem (MAUP) is ever-present and arises due to arbitrary decisions about the scale of analysis and the grouping of spatial units.

In this research, as with any ecological-level analysis, the Modifiable Areal Unit Problem (MAUP) emerges from arbitrary decisions regarding the scale of analysis and the grouping of spatial units. The MAUP consists of the scale effect, which concerns how the size of areal units used in an analysis can impact results, and the zonation effect, which refers to how changing the boundaries between spatial units of the approximate same size and number can change quantitative results (A S Fotheringham and Wong 1991; Dark and Bram 2007; M.-P. Parenteau and Sawada 2011). In the context of this study, the selection of areal units (neighborhood boundaries) affects the observed relationships between COVID-19 rates and the determinants. For instance, if the neighborhood borders were drawn or grouped larger or smaller, it will lead to different results for all models presented. As such, these findings are relevant only to the neighbourhood scale and Ottawa. Each study that was reviewed had different determinants and different overall relationships between determinants and COVID measures. There is no one-size fits all model. This limitation is acceptable because the neighbourhood scale is a scale at which Ottawa Health and the City of Ottawa use interventions for positive health outcomes. Hence, it was deemed justified to use this imposed neighbourhood boundaries over, for example, census tracts – not to mention the public data is only available at the neighbourhood scale.

Another limitation is the Uncertain Point Observation Problem (UPOP), which poses challenges in accurately determining the precise location of COVID-19 transmission cases, based on uncertain or incomplete information. In the context of COVID-19 studies, UPOP can lead to inconsistencies between the reported location of cases and the actual places where infections occur (Robertson and Feick 2018). This discrepancy arises because people may contract the virus in different locations from where they live. UPOP can introduce spatial biases in the data, potentially affecting the identification of high-transmission areas and the effectiveness of targeted public health interventions. For example, a neighbourhood with high COVID-19 rates would arise due to both local within neighbourhood contagions and remote out of neighbourhood contagions. All rates are determined based on where the cases live.

An important limitation to consider in this study relates to the temporal disparity between the socioeconomic data and the COVID-19 data. The socioeconomic variables utilized in this analysis were collected in 2016, while the COVID-19 data covered the period from July 2020 to December 2021. Therefore, it is necessary to exercise caution when interpreting the results. For more precise insights into the relationship between pre-existing socioeconomic determinants and COVID-19 outcomes, future research should aim to access more recent socioeconomic data.

This study represents one spatial perspective of COVID-19 analysis in Ottawa. While our current analysis provides insights into the determinants of COVID-19, it's important to note that there are several unexplored opportunities for future research that can further

broaden the scope of our understanding. For instance, future studies could consider conducting temporal comparisons across different seasons, such as summer and spring versus winter and fall and assess model transferability across time. Additionally, two regression models were created for neighborhoods displaying different patterns of spatial autocorrelation in COVID-19 rates. The first model will concentrate on neighborhoods with positive spatial autocorrelation, exhibiting high-high COVID-19 rates, while the second model will target neighborhoods with positive spatial autocorrelation but with low-low COVID-19 rates. There are numerous other analyses that could be done with this dataset, and our results represent one spatial perspective of COVID-19 analysis in Ottawa; there remain several unexplored opportunities to broaden the scope of research.

2. Conclusion

At the outset of this research the following questions were posed:

1. What is the spatial structure of COVID-19 in Ottawa and where are significant clusters and outliers of COVID-19 rates in Ottawa?
2. Using a top-down approach to modelling COVID-19 rates based on literature-derived socioeconomic health determinants, can local and global regression modelling accurately predict the pattern of COVID-19 cumulative rates in Ottawa?
3. Would a bottom-up Random Forest model find the same health determinants as important when compared to the regression modelling? Would RF provide a more accurate model?

Regarding the first research question, the COVID-19 pattern was clearly spatially autocorrelation and that was consistent with other studies reviewed in the literature review. Locally, there were numerous clustering of high and low COVID-19 rates. This highlighted some questionable decisions made at the provincial level as one neighbourhood that was prioritized for COVID-19 was situated well within a low-low cluster and was not a hot-spot. The characteristics of the cold spot neighbourhoods still require some explanation but certainly their characteristics would be useful for further research.

Regarding the second research question of identifying and modelling the COVID-19 rates using literature-identified health determinants, the study's findings underscore that some socioeconomic determinants contribute positively and some negatively to neighborhood COVID-19 rates. Top-down spatial statistical regression models highlight the significance of factors like educational attainment, age, and transportation modes in influencing these rates.

Regarding the last research question, using a bottom-up approach to identify potential health determinants, the Random Forest model revealed a largely different set of potential determinants than what was suggested by the literature. Specifically, neighborhoods with large numbers of Middle Eastern populations, income, migrant status, languages spoken, and marital status are additional that are important for explaining COVID-19 across the city. However, unraveling the mechanisms by which the identified potential health determinants influence health outcomes at the neighborhood level involves a degree of conjecture. It relies on explaining how individuals sharing common socioeconomic or

cultural traits might increase exposure or not adhere to public health advice, ultimately contributing to an increase in COVID-19 rates at the ecological level. Together, these diverse determinants collectively shape the intricate landscape of COVID-19 prevalence in Ottawa.

The overall differences in the important determinants selected from the top-down vs. bottom-up approaches were very different. The differences in determinants between approaches suggest, particularly in the case of Random Forest (RF), that there may indeed be confounding variables that we did not include and that would have well defined pathways to the COVID-19 patterns. This is certainly an area for future research. In addition, only 80% of the training data was used in the Random Forest model analysis to maintain some independent data for accuracy assessment. If all the data had been used in the RF analysis, then the important variables could have differed.

Despite the differences in determinants between our two approaches, the general approach taken in this thesis holds potential implications beyond Ottawa. By showcasing the applicability of these analytical techniques to regions with similar datasets, this research offers a framework for gaining insightful understanding COVID-19 spatial patterns on a broader scale. Policymakers and public health officials in other areas facing similar challenges can benefit from adopting similar methodologies to better plan their responses and implement targeted measures.

Chapter 6: References

Ager, D. E. 2005. “French Cultural, Languages and Telecommunications Policy Towards Sub-Saharan Africa.” *Modern & Contemporary France* 13 (1): 57–69. https://doi.org/10.1080/0963948052000341222.

Akhtar-Danesh, Noori, Andrea Baumann, Mary Crea-Arsenio, and Valentina Antonipillai. 2022. “COVID-19 Excess Mortality among Long-Term Care Residents in Ontario, Canada.” Edited by Gabriel Hoh Teck Ling. *PLOS ONE* 17 (1): e0262807. https://doi.org/10.1371/journal.pone.0262807.

Ali, Mohammad, Pierre Goovaerts, Nushrat Nazia, M Zahirul Haq, Mohammad Yunus, and Michael Emch. 2006. “[No Title Found].” *International Journal of Health Geographics* 5 (1): 45. https://doi.org/10.1186/1476-072X-5-45.

Allington, Daniel, Siobhan McAndrew, Vivienne Moxham-Hall, and Bobby Duffy. 2023. “Coronavirus Conspiracy Suspicions, General Vaccine Attitudes, Trust and Coronavirus Information Source as Predictors of Vaccine Hesitancy among UK Residents during the COVID-19 Pandemic.” *Psychological Medicine* 53 (1): 236–47. https://doi.org/10.1017/S0033291721001434.

Anselin, Luc. 2010. “Local Indicators of Spatial Association-LISA.” *Geographical Analysis* 27 (2): 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Banu, Shahera, Wenbiao Hu, Cameron Hurst, Yuming Guo, Mohammad Zahirul Islam, and Shilu Tong. 2012. “Space-Time Clusters of Dengue Fever in Bangladesh: **Space-Time Clusters of Dengue Fever**.” *Tropical Medicine & International Health* 17 (9): 1086–91. https://doi.org/10.1111/j.1365-3156.2012.03038.x.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bergstra, James, and Yoshua Bengio. 2012. “Random Search for Hyper-Parameter Optimization.” *Journal of Machine Learning Research* 13 (10): 281–305.

Bhadra, Arunava, Arindam Mukherjee, and Kabita Sarkar. 2021. “Impact of Population Density on Covid-19 Infected and Mortality Rate in India.” *Modeling Earth Systems and Environment* 7 (1): 623–29. https://doi.org/10.1007/s40808-020-00984-7.

Bidanset, Paul E., John R. Lombard, Peadar Davis, Michael McCord, and William J. McCluskey. 2017. “Further Evaluating the Impact of Kernel and Bandwidth Specifications of Geographically Weighted Regression on the Equity and Uniformity of Mass Appraisal Models.” In *Advances in Automated Valuation Modeling*, edited by Maurizio d’Amato and Tom Kauko, 86:191–99. Studies in Systems, Decision and Control. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-49746-4_11.

Bivand, Roger. 2022. “R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data.” *Geographical Analysis* 54 (3): 488–518. https://doi.org/10.1111/gean.12319.

Bivand, Roger, and Nicholas Lewin-Koh. 2016. “Maptools: Tools for Reading and Handling Spatial Objects.” https://CRAN.R-project.org/package=maptools.

Bivand, Roger, and Gianfranco Piras. 2015. “Comparing Implementations of Estimation Methods for Spatial Econometrics.” *Journal of Statistical Software* 63 (18). https://doi.org/10.18637/jss.v063.i18.

Bivand, Roger S., Edzer Pebesma, and Virgilio Gómez-Rubio. 2013. *Applied Spatial Data Analysis with R*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7618-4.

Breiman, Leo. 2001a. "[No Title Found]." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

———. 2001b. "[No Title Found]." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Brunsdon, C., S. Fotheringham, and M. Charlton. 1998. "Geographically Weighted Regression." *Journal of the Royal Statistical Society: Series D (The Statistician)* 47 (3): 431–43. https://doi.org/10.1111/1467-9884.00145.

Burton, Alexander L. 2021. "OLS (Linear) Regression." In *The Encyclopedia of Research Methods in Criminology and Criminal Justice*, edited by J.C. Barnes and David R. Forde, 1st ed., 509–14. Wiley. https://doi.org/10.1002/9781119111931.ch104.

Castro, R. R., R. S. C. Santos, G. J. B. Sousa, Y. T. Pinheiro, R. R. I. M. Martins, M. L. D. Pereira, and R. A. R. Silva. 2021. "Spatial Dynamics of the COVID-19 Pandemic in Brazil." *Epidemiology and Infection* 149: e60. https://doi.org/10.1017/S0950268821000479.

Chen, Shi-Yi, Zhe Feng, and Xiaolian Yi. 2017. "A General Introduction to Adjustment for Multiple Comparisons." *Journal of Thoracic Disease* 9 (6): 1725–29. https://doi.org/10.21037/jtd.2017.05.34.

Clark, Eva, Karla Fredricks, Laila Woc-Colburn, Maria Elena Bottazzi, and Jill Weatherhead. 2020. "Disproportionate Impact of the COVID-19 Pandemic on Immigrant Communities in the United States." Edited by Victoria J. Brookes. *PLOS Neglected Tropical Diseases* 14 (7): e0008484. https://doi.org/10.1371/journal.pntd.0008484.

Cordes, Jack, and Marcia C. Castro. 2020. "Spatial Analysis of COVID-19 Clusters and Contextual Factors in New York City." *Spatial and Spatio-Temporal Epidemiology* 34 (August): 100355. https://doi.org/10.1016/j.sste.2020.100355.

Dark, Shawna J., and Danielle Bram. 2007. "The Modifiable Areal Unit Problem (MAUP) in Physical Geography." *Progress in Physical Geography*. https://doi.org/10.1177/0309133307083294.

Darques, Regis, Julie Trottier, Raphael Gaudin, and Nassim Ait-Mouheb. 2022. "Clustering and Mapping the First COVID-19 Outbreak in France." *BMC Public Health* 22 (1): 1279. https://doi.org/10.1186/s12889-022-13537-7.

Demenech, Lauro Miranda, Samuel De Carvalho Dumith, Maria Eduarda Centena Duarte Vieira, and Lucas Neiva-Silva. 2020. "Desigualdade Econômica e Risco de Infecção e Morte Por COVID-19 No Brasil." *Revista Brasileira de Epidemiologia* 23: e200095. https://doi.org/10.1590/1980-549720200095.

Draper, Norman R., and Harry Smith. 1998. *Applied Regression Analysis*. 1st ed. Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/9781118625590.

Elgar, Frank J., Anna Stefaniak, and Michael J.A. Wohl. 2020. "The Trouble with Trust: Time-Series Analysis of Social Capital, Income Inequality, and COVID-19 Deaths in 84 Countries." *Social Science & Medicine* 263 (October): 113365. https://doi.org/10.1016/j.socscimed.2020.113365.

ESRI. 2023a. "Exploratory Regression (Spatial Statistics)—ArcGIS Pro | Documentation." 2023. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/exploratory-regression.htm.

———. 2023b. “Forest-Based Classification and Regression.” ESRI. 2023. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/forestbasedclassificationregression.htm.
———. 2023a. “How Forest-Based and Boosted Classification and Regression Works—ArcGIS Pro | Documentation.” 2023a. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-forest-works.htm.
Fotheringham, A S, and D W S Wong. 1991. “The Modifiable Areal Unit Problem in Multivariate Statistical Analysis.” *Environment and Planning A: Economy and Space* 23 (7): 1025–44. https://doi.org/10.1068/a231025.
Fotheringham, A. Stewart, Wenbai Yang, and Wei Kang. 2017a. “Multiscale Geographically Weighted Regression (MGWR).” *Annals of the American Association of Geographers* 107 (6): 1247–65. https://doi.org/10.1080/24694452.2017.1352480.
———. 2017b. “Multiscale Geographically Weighted Regression (MGWR).” *Annals of the American Association of Geographers* 107 (6): 1247–65. https://doi.org/10.1080/24694452.2017.1352480.
Ghosh, Pritam, and Alfredo Cartone. 2020. “A Spatio-temporal Analysis of COVID-19 Outbreak in Italy.” *Regional Science Policy & Practice* 12 (6): 1047–62. https://doi.org/10.1111/rsp3.12376.
Gollini, Isabella, Binbin Lu, Martin Charlton, Christopher Brunsdon, and Paul Harris. 2015a. “**GWmodel** : An *R* Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models.” *Journal of Statistical Software* 63 (17). https://doi.org/10.18637/jss.v063.i17.
———. 2015b. “GWmodel : An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models.” *Journal of Statistical Software* 63 (17). https://doi.org/10.18637/jss.v063.i17.
Government of Canada, Statistics Canada. 2021. “Impacts of the COVID-19 Pandemic in Nursing and Residential Care Facilities in Canada.” June 10, 2021. https://www150.statcan.gc.ca/n1/pub/45-28-0001/2021001/article/00025-eng.htm.
Guo, Yan-Rong, Qing-Dong Cao, Zhong-Si Hong, Yuan-Yang Tan, Shou-Deng Chen, Hong-Jun Jin, Kai-Sen Tan, De-Yun Wang, and Yan Yan. 2020. “The Origin, Transmission and Clinical Therapies on Coronavirus Disease 2019 (COVID-19) Outbreak – an Update on the Status.” *Military Medical Research* 7 (1): 11. https://doi.org/10.1186/s40779-020-00240-0.
Hamidi, Shima, Sadegh Sabouri, and Reid Ewing. 2020. “Does Density Aggravate the COVID-19 Pandemic?: Early Findings and Lessons for Planners.” *Journal of the American Planning Association* 86 (4): 495–509. https://doi.org/10.1080/01944363.2020.1777891.
Han, Yi, Lan Yang, Kun Jia, Jie Li, Siyuan Feng, Wei Chen, Wenwu Zhao, and Paulo Pereira. 2021. “Spatial Distribution Characteristics of the COVID-19 Pandemic in Beijing and Its Relationship with Environmental Factors.” *Science of The Total Environment* 761 (March): 144257. https://doi.org/10.1016/j.scitotenv.2020.144257.
Hawkins, R.B., E.J. Charles, and J.H. Mehaffey. 2020. “Socio-Economic Status and COVID-19–Related Cases and Fatalities.” *Public Health* 189 (December): 129–34. https://doi.org/10.1016/j.puhe.2020.09.016.

Hoare, Jake. 2018. “How Is Variable Importance Calculated for a Random Forest?” *Displayr* (blog). July 30, 2018. https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/.
Hu, Lan, Yongwan Chun, and Daniel A. Griffith. 2022. “Incorporating Spatial Autocorrelation into House Sale Price Prediction Using Random Forest Model.” *Transactions in GIS* 26 (5): 2123–44. https://doi.org/10.1111/tgis.12931.
James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7.
Kulldorf, M. 2023. “SaTScan - Software for the Spatial, Temporal, and Space-Time Scan Statistics.” 2023. https://www.satscan.org/.
Kulldorff, Martin. 1997. “A Spatial Scan Statistic.” *Communications in Statistics - Theory and Methods* 26 (6): 1481–96. https://doi.org/10.1080/03610929708831995.
Kuznetsov, Andrey, and Veronika Sadovskaya. 2021. “Spatial Variation and Hotspot Detection of COVID-19 Cases in Kazakhstan, 2020.” *Spatial and Spatio-Temporal Epidemiology* 39 (November): 100430. https://doi.org/10.1016/j.sste.2021.100430.
Lee, Sang-Il. 2001. “Developing a Bivariate Spatial Association Measure: An Integration of Pearson’s r and Moran’s I.” *Journal of Geographical Systems* 3 (4): 369–85. https://doi.org/10.1007/s101090100064.
———. 2004a. “A Generalized Significance Testing Method for Global Measures of Spatial Association: An Extension of the Mantel Test.” *Environment and Planning A: Economy and Space* 36 (9): 1687–1703. https://doi.org/10.1068/a34143.
———. 2004b. “A Generalized Significance Testing Method for Global Measures of Spatial Association: An Extension of the Mantel Test.” *Environment and Planning A: Economy and Space* 36 (9): 1687–1703. https://doi.org/10.1068/a34143.
LeSage, James P. 2008. “An Introduction to Spatial Econometrics.” *Revue d’économie Industrielle*, no. 123 (September): 19–44. https://doi.org/10.4000/rei.3887.
LeSage, James, and Robert Kelley Pace. 2009. *Introduction to Spatial Econometrics*. 0 ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781420064254.
Li, Ziqi, A. Stewart Fotheringham, Taylor M. Oshan, and Levi John Wolf. 2020. “Measuring Bandwidth Uncertainty in Multiscale Geographically Weighted Regression Using Akaike Weights.” *Annals of the American Association of Geographers* 110 (5): 1500–1520. https://doi.org/10.1080/24694452.2019.1704680.
Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2 (3): 18–22.
Lombard, John R., Eliahu Stern, and Graham Clarke, eds. 2016. *Applied Spatial Modelling and Planning*. 0 ed. Routledge. https://doi.org/10.4324/9781315683621.
Mansfield, Edward R., and Billy P. Helms. 1982. “Detecting Multicollinearity.” *The American Statistician* 36 (3a): 158–60. https://doi.org/10.1080/00031305.1982.10482818.
Manski, Charles F. 1993. “Identification of Endogenous Social Effects: The Reflection Problem.” *The Review of Economic Studies* 60 (3): 531. https://doi.org/10.2307/2298123.
Mansour, Shawky, Abdullah Al Kindi, Alkhattab Al-Said, Adham Al-Said, and Peter Atkinson. 2021. “Sociodemographic Determinants of COVID-19 Incidence Rates in Oman: Geospatial Modelling Using Multiscale Geographically Weighted

Regression (MGWR)." *Sustainable Cities and Society* 65 (February): 102627. https://doi.org/10.1016/j.scs.2020.102627.
Menard, Scott. 2002. *Applied Logistic Regression Analysis*. SAGE.
Moran, P. A. P. 1950. "NOTES ON CONTINUOUS STOCHASTIC PHENOMENA." *Biometrika* 37 (1–2): 17–23. https://doi.org/10.1093/biomet/37.1-2.17.
Ottawa Neighborhood Study. n.d. "Ottawa Neighbourhood Study | Our Aim Is to Better Understand the Physical and Social Pathways through Which Neighbourhoods in Ottawa Affect Health." Accessed December 3, 2023. https://www.neighbourhoodstudy.ca/.
Pamidimukkala, Apurva, and Sharareh Kermanshachi. 2021. "Impact of Covid-19 on Field and Office Workforce in Construction Industry." *Project Leadership and Society* 2 (December): 100018. https://doi.org/10.1016/j.plas.2021.100018.
Parenteau, M. P., M. Sawada, Elizabeth A. Kristjansson, Melissa Calhoun, Stephanie Leclair, Ronald Labonté, V. Runnels, Anna-Lena Musiol, and Sam Herold. 2008. "Development of Neighborhoods to Measure Spatial Indicators of Health." *Urisa Journal* 20: 43–55.
Parenteau, Marie-Pierre, and Michael C. Sawada. 2011. "The Modifiable Areal Unit Problem (MAUP) in the Relationship between Exposure to NO 2 and Respiratory Health." *International Journal of Health Geographics*. https://doi.org/10.1186/1476-072X-10-58.
Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science: With Applications in R*. 1st ed. New York: Chapman and Hall/CRC. https://www.taylorfrancis.com/books/9780429459016.
Péterfi, Anna, Ágota Mészáros, Zsófia Szarvas, Melinda Pénzes, Mónika Fekete, Ágnes Fehér, Andrea Lehoczki, Tamás Csípő, and Vince Fazekas-Pongor. 2022. "Comorbidities and Increased Mortality of COVID-19 among the Elderly: A Systematic Review." *Physiology International* 109 (2): 163–76. https://doi.org/10.1556/2060.2022.00206.
Rainham, Daniel, Ian McDowell, Daniel Krewski, and Mike Sawada. 2010. "Conceptualizing the Healthscape: Contributions of Time Geography, Location Technologies and Spatial Ecology to Place and Health Research." *Social Science & Medicine* 70 (5): 668–76. https://doi.org/10.1016/j.socscimed.2009.10.035.
Rezaeian, M., G. Dunn, S. St Leger, and L. Appleby. 2007. "Geographical Epidemiology, Spatial Analysis and Geographical Information Systems: A Multidisciplinary Glossary." *Journal of Epidemiology & Community Health* 61 (2): 98–102. https://doi.org/10.1136/jech.2005.043117.
Robertson, Colin, and Rob Feick. 2018. "Inference and Analysis across Spatial Supports in the Big Data Era: Uncertain Point Observations and Geographic Contexts." *Transactions in GIS* 22 (2): 455–76. https://doi.org/10.1111/tgis.12321.
Rocklöv, Joacim, and Henrik Sjödin. 2020. "High Population Densities Catalyse the Spread of COVID-19." *Journal of Travel Medicine* 27 (3): taaa038. https://doi.org/10.1093/jtm/taaa038.
Rossiter, D G. 2022. "Geographically Weighted Models." Cornell University's Soil & Crop Sciences Section and Nanjing Normal University's Geographic Sciences Department. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.css.cornell.edu/faculty/dgr2/_static/files/ov/GWR_Handout.pdf.

Sallam, Malik, Deema Dababseh, Huda Eid, Kholoud Al-Mahzoum, Ayat Al-Haidar, Duaa Taim, Alaa Yaseen, Nidaa A. Ababneh, Faris G. Bakri, and Azmi Mahafzah. 2021. "High Rates of COVID-19 Vaccine Hesitancy and Its Association with Conspiracy Beliefs: A Study in Jordan and Kuwait among Other Arab Countries." *Vaccines* 9 (1): 42. https://doi.org/10.3390/vaccines9010042.

Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52 (3–4): 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Shi, Xuerui, Gabriel Hoh Teck Ling, Pau Chung Leng, Noradila Rusli, and Ak Mohd Rafiq Ak Matusin. 2023. "Associations between Institutional-Social-Ecological Factors and COVID -19 Case-Fatality: Evidence from 134 Countries Using Multiscale Geographically Weighted Regression (MGWR)." *One Health* 16 (June): 100551. https://doi.org/10.1016/j.onehlt.2023.100551.

Siljander, Mika, Ruut Uusitalo, Petri Pellikka, Sanna Isosomppi, and Olli Vapalahti. 2022. "Spatiotemporal Clustering Patterns and Sociodemographic Determinants of COVID-19 (SARS-CoV-2) Infections in Helsinki, Finland." *Spatial and Spatio-Temporal Epidemiology* 41 (June): 100493. https://doi.org/10.1016/j.sste.2022.100493.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8 (1): 25. https://doi.org/10.1186/1471-2105-8-25.

Suthaharan, Shan. 2016. "Decision Tree Learning." In *Machine Learning Models and Algorithms for Big Data Classification*, by Shan Suthaharan, 36:237–69. Integrated Series in Information Systems. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4899-7641-3_10.

Tang, Ian W., Verónica M. Vieira, and Eric Shearer. 2022. "Effect of Socioeconomic Factors during the Early COVID-19 Pandemic: A Spatial Analysis." *BMC Public Health* 22 (1): 1212. https://doi.org/10.1186/s12889-022-13618-7.

Terrell, George R., and David W. Scott. 1992. "Variable Kernel Density Estimation." *The Annals of Statistics* 20 (3). https://doi.org/10.1214/aos/1176348768.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (June): 234. https://doi.org/10.2307/143141.

Valero-Martínez, Carla, Christopher Martínez-Rivera, Jenny Zhen-Duan, Marie Fukuda, and Margarita Alegría. 2023. "Attitudes toward COVID-19 Vaccine Uptake: A Qualitative Study of Mostly Immigrant Racial/Ethnic Minority Older Adults." *Geriatrics* 8 (1): 17. https://doi.org/10.3390/geriatrics8010017.

Vilches, Thomas N., Shokoofeh Nourbakhsh, Kevin Zhang, Lyndon Juden-Kelly, Lauren E. Cipriano, Joanne M. Langley, Pratha Sah, Alison P. Galvani, and Seyed M. Moghadas. 2021. "Multifaceted Strategies for the Control of COVID-19 Outbreaks in Long-Term Care Facilities in Ontario, Canada." *Preventive Medicine* 148 (July): 106564. https://doi.org/10.1016/j.ypmed.2021.106564.

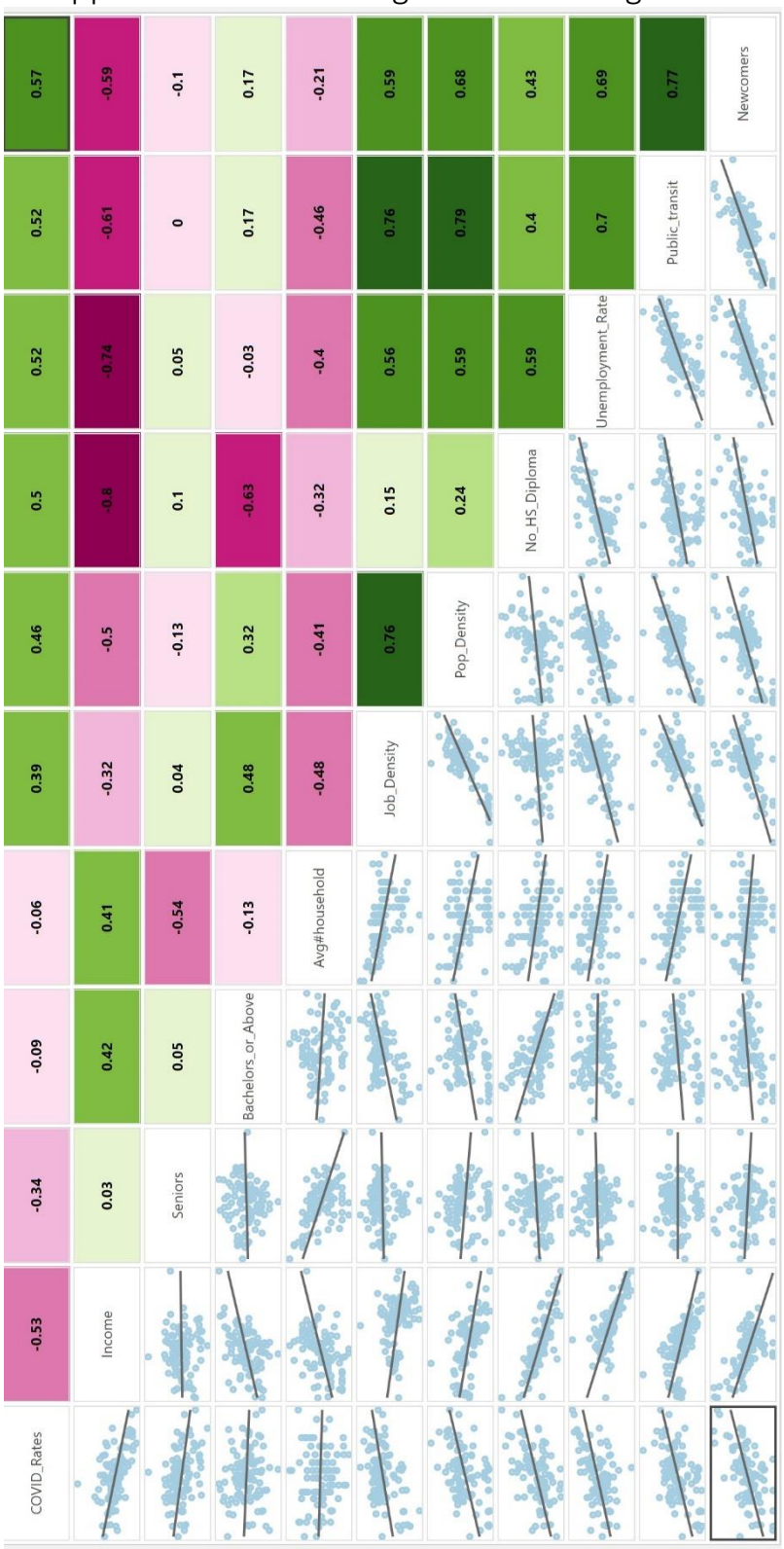
Vilinová, Katarína, and Lucia Petrikovičová. 2023. "Spatial Autocorrelation of COVID-19 in Slovakia." *Tropical Medicine and Infectious Disease* 8 (6): 298. https://doi.org/10.3390/tropicalmed8060298.

Wheeler, David C. 2014. "Geographically Weighted Regression." In *Handbook of Regional Science*, edited by Manfred M. Fischer and Peter Nijkamp, 1435–59. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23430-9_77.

Wichura, Michael J. 1988. “Algorithm AS 241: The Percentage Points of the Normal Distribution.” *Applied Statistics* 37 (3): 477. https://doi.org/10.2307/2347330.
Wong, David W. S., and Yun Li. 2020. “Spreading of COVID-19: Density Matters.” Edited by Bing Xue. *PLOS ONE* 15 (12): e0242398. https://doi.org/10.1371/journal.pone.0242398.
Wu, Jianhong, Biao Tang, Nicola Luigi Bragazzi, Kyeongah Nah, and Zachary McCarthy. 2020. “Quantifying the Role of Social Distancing, Personal Protection and Case Detection in Mitigating COVID-19 Outbreak in Ontario, Canada.” *Journal of Mathematics in Industry* 10 (1): 15. https://doi.org/10.1186/s13362-020-00083-3.
Zarbo, Cristina, Valentina Candini, Clarissa Ferrari, Miriam d’Addazio, Gemma Calamandrei, Fabrizio Starace, Marta Caserotti, et al. 2022. “COVID-19 Vaccine Hesitancy in Italy: Predictors of Acceptance, Fence Sitting and Refusal of the COVID-19 Vaccination.” *Frontiers in Public Health* 10 (April): 873098. https://doi.org/10.3389/fpubh.2022.873098.

Appendices

Appendix 1: This is a larger version of Figure 4.



Appendix 2: The custom Local Moran's I function used throughout the analysis.

```
localmoran.mc <- function (invector, inadjmatrix, mlvar=TRUE,nsim =
                     99){
  require(spdep)

  # z-score the data. See localmoran() mlvar for
  # explanation of this choice of variance
  if (!mlvar)
    z = scale(invector)[, 1]
  else
    z = (invector - mean(invector)) / (sd(invector) *
                                 sqrt((length(invector) - 1) / length(invector)))

  # compute the observed local moran's I
  obs = mapply(
    function(X, Y, V, v) {
      V * sum(v[X] * Y)
    },
    X = inadjmatrix$neighbours,
    Y = inadjmatrix$weights,
    V = as.list(z),
    MoreArgs = list(v = z)
  )
  # Create a list of ids for each row so that the current observation
  idx = as.list(1:length(invector))
  # repeat the calculation of local I nsim times,
  # for each row calculate Li by shuffling all n-1 data values
  refd = replicate(
    nsim,
    mapply(
      function(X, Y, V, v, rid) {
        v[-rid] = sample(v[-rid])
        #gx[,rid]<<-v
        V * sum(v[X] * Y)
      },
      X = inadjmatrix$neighbours,
      Y = inadjmatrix$weights,
      V = as.list(z),
      rid = idx,
      MoreArgs = list(v = z)
    )
  )
  res = cbind(obs, refd)
  less = rowSums(sweep(res, 1, res[, 1], "<=")) / (nsim + 1)
```

```
greater = rowSums(sweep(res, 1, res[, 1], ">=")) / (nsim + 1)
ponesided = apply(cbind(less, greater), 1, min)
resscale=t(scale(t(res)))
both = rowSums(
  sweep(abs(resscale),
        1,
        abs(resscale)[, 1],
        ">=")
) / (nsim + 1)
fdrponesided = p.adjustSP(
  ponesided,
  inadjmatrix$neighbours,
  method = "fdr")

fdrboth = p.adjustSP(
  both,
  inadjmatrix$neighbours,
  method = "fdr")

cDV = z
c_mI = as.vector(lag.listw(inadjmatrix, z))

cDVmean = mean(cDV)
miMean = mean(c_mI)
quadrant = vector(mode = "character", length = length(ponesided))
quadrant[cDV > cDVmean & c_mI > miMean] = "HH"
quadrant[cDV < cDVmean & c_mI > miMean] = "LH"
quadrant[cDV > cDVmean & c_mI < miMean] = "HL"
quadrant[cDV < cDVmean & c_mI < miMean] = "LL"
# set a statistical significance level for the local Moran's
signif = 0.05
# places non-significant Moran's in the category "5"
quadrant[ponesided > signif] = "NOTSIG"

quadrant=factor(quadrant,levels=c("HH", "LL", "LH", "HL",
                          "NOTSIG"),ordered=T)

# Create list of results
return(list(
  results = data.frame(
    Ii = res[, 1],
    one.sided = ponesided,
    FDRone.sided = fdrponesided,
    two.sided = both,
    FDRtwo.sided = fdrboth,
    quadrant = quadrant
  ),
  raw = res,
```

```
   moran.scatter = list(x=cDV,y=c_mI)
  ))
}
```

Appendix 3: List of all variables included in the Random Forest Model.

Percentage residents within 15 min walking distance to recreation center
Average distance to nearest community recreation centre (km)
Average distance to nearest 3 community recreation centres (km)
Number of community recreation centres (with 50m buffer) per 1000 people
Percentage Tree Canopy
Job density (Number jobs per km2)
Percentage of jobs that are full-time
Percentage of jobs in part-time
Percentage of jobs in agriculture and resource extraction
Percentage of jobs in construction and utilities
Percentage of jobs in manufacturing and transportation
Percentage of jobs in information, education, and science
Percentage of jobs in finance, real estate, and enterprises
Percentage of jobs in health care and social assistance
Percentage of jobs in retail, accommodation, and recreation
Percentage of jobs in other services
Percentage residents within 15 min walking distance of CHC/CRC
Average distance to nearest CHC/CRC (km)
Average distance to nearest 3 French elementary schools (km)
Percentage residents within 15 min walking distance of elementary school
Average distance to nearest 3 French high schools (km)
Percentage residents within 15 min walking distance of high school
Total road network length (km)
Road network density (km/km2)
Road network length per capita (km/1000 residents)
Perceived Walkability
Average Number of public transit stops within 600m
Average Number of public transit routes stopping within 600m
Number of public transit stops per 1000 people
Number of public transit stops per km2
Average distance to nearest 3 coffee shops (km)
Number of coffee shops (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to coffee shops
Average distance to nearest 3 convenience stores (km)
Number of convenience stores (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to convenience stores
Average distance to nearest 3 fast food outlets (km)
Number of fast food outlets (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to fast food outlets
Average distance to nearest 3 grocery stores (km)

Number of grocery stores (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to a grocery store
Supermarkets: Average distance to nearest 3 supermarkets (km)
Supermarkets: Number of supermarkets (with 50m buffer) per 1000 people
Supermarkets: Percentage residents within 15 min walking distance
Average distance to nearest 3 nightlife venues (km)
Number of nightlife venues (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to nightlife venues
Average distance to nearest 3 pharmacies (km)
Number of pharmacies (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to pharmacies
Pharmacies: Average distance to nearest 3 pharmacies (km)
Pharmacies: Number of pharmacies (with 50m buffer) per 1000 people
Pharmacies: Percentage residents within 15 min walking distance
Average distance to nearest 3 restaurants (km)
Number of restaurants (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to restaurants
Average distance to nearest 3 specialty food stores (km)
Number of specialty food stores (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to food stores
Average distance to nearest 3 sweet and treats (km)
Number of sweet and treats (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to candy stores
Average distance to nearest optometry clinic (km)
Number of optometry clinics (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to optometry clinic
Number of optometrists (with 50m buffer) per 1000 people
Socioeconomic Index
Average distance to nearest 3 outdoor summer sport areas (km)
Number of outdoor summer sport areas (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to summer sport areas
Average distance to nearest 3 cool-off areas (km)
Number of cool-off areas (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to cool-off areas
Average distance to nearest 3 winter outdoor recreation areas (km)
Number of winter outdoor recreation areas (with 50m buffer) per 1000 people
Pedestrian network density (km/km2)
Pedestrian network length per capita (km/1000 residents)
Total cycling network length (km)
Cycling network density (km/km2)
Cycling network length per capita (km/1000 residents)
Pedestrian collisions (with 50m buffer) 2014-2018
Total collisions per 1km road length (with 50m buffer) 2014-2018
Average distance to the nearest bike repair station (km)
Average distance to nearest 3 bike repair stations (km)
Number of bike repair stations per 1000 people
Percentage residents within 15 min walking distance to bike repair
Average distance to nearest 1 publicly accessible computer site (km)

Average distance to nearest 3 publicly accessible computer sites (km)
Number of publicly accessible computer sites (with 50m buffer) per 1000 people
Trails density (km/km2)
Trails length per capita (km/1000 residents)
Percentage residents within 15 min walking distance of a public toilet
Number of public toilets
Percentage residents not within 7min driving distance of a fire station
Neighbourhood area (km2)
All Crimes against the person (total Number of, per 1,000 residents)
All Crimes against the property (total Number of, per 1,000 residents)
All "Other" criminal code offenses (total Number of, per 1,000 residents)
All Crimes (total Number of, per 1,000 residents)
Average Number of annual shootings 2016-2019
Average Number of Annual Bicycle Thefts (2015-2019)
Percentage residents within 15 min walking distance to public library
Average distance to nearest Ottawa Public Library (km)
Average distance to nearest 3 English elementary schools (km)
Percentage residents within 15 min walking distance to English elementary school
Average distance to nearest 3 English high schools (km)
Percentage residents within 15 min walking distance to English high school
Percentage of children low in 1 or more EDI domains
Percentage of children vulnerable in communication skills and general knowledge
Percentage of children vulnerable in emotional maturity
Percentage of children vulnerable in language and cognitive development
Percentage of children vulnerable in physical health and well-being
Percentage of children vulnerable in social competence
Percentage children who reported French as a first language spoken
Percentage children who speak a non-official language
Percentage children who are in French Immersion
Average distance to nearest 3 dental clinics (km)
Number of dental clinics (with 50m buffer) per 1000 people
Percentage residents within 15 min walking distance to dental clinic
Number of dentists (with 50m buffer) per 1000 people
Percentage Owner households
Percentage Renter households
Percentage Unsuitable housing
Percentage Major repairs needed
Percentage Unaffordable housing (all households)
Percentage Unaffordable housing (tenant households)
Percentage Tenants in subsidized housing
Median monthly rent cost ($)
Percentage Moved in the last year
Percentage Moved in the last 5 years
Percentage Built in the 50's or earlier
Percentage Built in the 60's & 70's
Percentage Built in the 80's or 90's
Percentage Built in the 2000's
Median resident income (after taxes)

Median resident total income (before taxes)
Median resident employment income
Median amount of resident government transfer
Low income prevalence (LIM-AT)
Low income prevalence (LIM-AT) among children (0-14)
Low income prevalence (LIM-AT) among youth (15-24)
Low income prevalence (LIM-AT) among adults (25-64)
Low income prevalence (LIM-AT) among seniors (65+)
Unemployment rate
Number Unemployed
Percentage Employed full time for full year
Percentage Self-employed
Percentage With no high school diploma
Percentage High school level
Percentage Above high school, below bachelor
Percentage Bachelor level or above
Percentage Drive to work (as driver or passenger)
Percentage Take public transit to work
Percentage Walk to work
Percentage Bike to work
Percentage With less than 15-minute commute to work
Percentage With 15-29 minute commute to work
Percentage With 30+ minute commute to work
Percentage Leaving for work before 7 am
Percentage Leaving for work between 7 am-8 am
Percentage Leaving for work between 8 am-9 am
Percentage Leaving for work after 9 am or before 5 am
Total population (2016)
Number Children (0-14)
Number Youth (15-24)
Number Adults (25-64)
Number Seniors (65+)
Percentage Children (0-14)
Percentage Youth (15-24)
Percentage Adults (25-64)
Percentage Seniors (65+)
Median age
Percentage Married/Common law
Percentage Single (never married)
Percentage Separated/Divorced/Widowed
Number Newcomers (2011-2016)
Number Newcomers (2006-2010)
Number Non-permanent residents
Percentage Newcomers (2011-2016)
Percentage Newcomers (2006-2010)
Percentage Non-permanent residents
Percentage First-generation immigrants
Percentage Second-generation immigrants

Percentage Indigenous People
Percentage Single-parent families
Average household size
Number People living alone
Percentage People living alone
Percentage Households with multiple families
Median household income (AT)
Percentage Who can speak neither English nor French
Percentage Who can speak English only
Percentage Who can speak French only
Percentage Who can speak both English and French
Median market income
Labour force participation rate
Percentage Racialized population
Percentage South Asian
Percentage Chinese
Percentage Black
Percentage Filipino
Percentage Latin American
Percentage Middle-Eastern
Percentage Southeast Asian
Percentage West Asian
Percentage Korean
Percentage Japanese
Percentage Other racialized population
Percentage Non-racialized population
Percentage Seniors (65+) living alone
Youth (15-25) unemployment rate
Seniors (65+) unemployment rate
Number advanced age (80+) seniors
Percentage of seniors who are recent immigrants
Percentage of seniors with no high school diploma
Percentage of seniors in the labour force
Number of census families
Average size of census families
Percentage one-census-family households
Percentage multiple-census-family households
Percentage non-census-family households
Percentage refugees
Percentage of seniors who speak neither English nor French
Low income prevalence (LIM-AT) among young children (0-5)
Population density (Number/km2)
Percentage residents within 15 min walking distance of a physician
Number of physicians in the neighborhood
Number of physicians (with 50m buffer) per 1000 people
Average drive time to the nearest 5 physicians (min)
Average walking distance to the nearest 5 physicians (km)
Population Density with square root transformation

Unemployment rate with log transformation
Percentage of newcomers with square root transformation
Job density with log transformation
Percentage of people with no high school diploma with square root transformation
Percentage of people with a bachelor's level degree or above with square root transformation
Percentage of people over 65 years old with log transformation
COVID-19 Cumulative rates with log transformation.